



**Population and family based studies of Consanguinity: Genetic
and Computational approaches**

Abdullah Mesut Erzurumluoğlu

*A dissertation submitted to the University of Bristol in accordance with the requirements for
award of degree of Doctor of Philosophy (PhD) in the Faculty of Medicine and Dentistry*

School of Social and Community Medicine

October 2015

****Redacted/revised version – signatures removed****

Word Count = ~65,000*

*Excluding preface, tables, footnotes, references and appendices

Abstract

Consanguinity is the union of closely related individuals – which can have genetic implications on the health of offspring(s). Consanguineous families with disorders have been extensively analysed by geneticists and this has led to the identification of many autosomal recessive disorder causal variants and genes. Two copies of the ‘inactivating’ or loss of function (LoF) allele are required to cause an autosomal recessive disorder, one inherited from the mother and the other from the father. In outbreeding populations these LoF alleles very rarely meet their counterpart (as it requires **both** parents to possess the allele), thus are passed down the generations silently – sometimes for millennia. However, consanguineous and/or endogamous offspring have elevated levels of homozygosity, which dramatically increases the probability of any allele to be in a homozygous (or more correctly autozygous) state. This increase in probability applies to LoF mutations also; and this elevation of levels of homozygosity is the main reason why extremely rare autosomal recessive disorders are usually only seen in populations where consanguinity (and/or endogamy) levels are high.

With the ever decreasing prices of DNA sequencing, whole-genome sequencing is becoming a reality for many laboratories. However, for now, whole-exome sequencing (WES) is the most feasible sequencing technique mostly due to cost factors. Combining the concepts of consanguinity and WES, the aim of this thesis was to identify ‘causal’ variants by analysing whole-exome data obtained from consanguineous families/individuals affected from autosomal recessive disorders such as Primary Ciliary Dyskinesia (PCD) and Autosomal Recessive Intellectual Disability (ARID). Using autozygosity mapping, a novel region located on chromosome 19 (p13.3) was identified to be associated with ARID (later sequenced by another research group and *ADAT3* was identified as the causal gene). Using WES, rare homozygous nonsense mutations p.E309* in *CCDC151* and p.R136* in *DNAAF3* were found to be causal of PCD. Other variants such as p.M263T in *MNS1*, p.R263* in *DNALI1*, p.G734fs in *HEATR2* and p.E328* in *LRRC48* have also been identified which may be causal of PCD but studies in this thesis remained inconclusive due to various reasons. Additionally, a rare missense mutation p.G300D in *CTSC* was found to be Papillon-Léfevre syndrome (PLS) causal. This latter finding illustrated the additional information that can be gained from WES data – which is discussed in the thesis.

Finding novel causal variants and gene functions can improve genetic counselling and lead to the identification of targets for preventive and/or curative medicine. In this respect, analysing consanguineous populations as a whole rather than ‘cherry picking’ families with disorders will have additional benefits and facilitate our understanding of the human genome – and this subject is also discussed in this thesis.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED:

DATE: 01/10/2015

Acknowledgements

I would like to begin by thanking every single one of my colleagues at the Bristol Genetic Epidemiology Laboratories (BGEL) for their academic, social and emotional help over the four years I have been there. Special mentions should go to my supervisors Dr. Santiago Rodriguez, Dr. Tom R. Gaunt and Prof. Ian Day, my desk mates (and colleagues) Dr. Hashem A. Shihab, Denis A. Baird, Tom G. Richardson, Dr. Jie Zheng and Dr. Chris Boustred, and colleagues Dr. John Kemp, Dr. Philip Guthrie and Dr. Osama Al-Ghamdi for their extra effort and time on transforming the academically 'naïve and inexperienced' Mesut that joined the BGEL in January 2012 to Dr. Erzurumluoğlu today.

I am indebted to my parents Ayla and (Dr.) Bayram Erzurumluoğlu, and siblings (Esat, Hasna and Nuran) for the sacrifices they have made; my friends and housemates for putting up with me; and most of all, God Almighty (The Most Gracious, The Most Merciful) for giving me the ability, strength and will power to overcome obstacles and get through hard times in academic, economic and social life.

Finally, I am grateful to the Medical Research Council (MRC) for the scholarship they have provided as it enabled me to concentrate solely on my research and therefore directly contributed to the findings published in this thesis.

“Kulları içinde ancak âlimler, Allah’ı gerektiği tarzda tazim ederler.”

Kuran-ı Kerîm, Fâtir, 28

“Among His servants/creation, only the scholars (those who have knowledge) truly fear and honour God.”

Holy Qur’an, Fatir, 28

Table of Contents

Chapter 1. Introduction and Literature Review	1
1.1. Organism to Genome to Genes	1
1.1.1. What is a gene?.....	4
1.1.2. Variation in a genome	8
1.2. Epidemiology and Genetics.....	15
1.2.1. Genetic Epidemiology	15
1.2.2. Terminology.....	17
1.2.3. Mendelian disorders.....	20
1.2.4. Past and present hypotheses on Mendelian disorders.....	24
1.2.5. Complex disorders.....	26
1.2.6. Past and present hypotheses on Complex disorders.....	30
1.3. Consanguinity and Genetic research.....	33
1.3.1. Consanguineous societies and genetic disease.....	36
1.3.2. Inbreeding depression in humans?	37
1.3.3. Historical perspective.....	45
1.3.4. Autozygosity	50
1.3.5. World-wide Consanguinity.....	55
1.4. Identifying the Genetic basis of human diseases.....	60
1.4.1. Traditional methods	61
1.4.2. Current methods	62
1.5. Detecting DNA sequence variation.....	63
1.5.1. Whole genome sequencing.....	65
1.5.2. Whole exome sequencing	66
1.5.3. Other methods.....	68

1.6.	DNA Sequencing technologies.....	71
1.6.1.	Historical background.....	71
1.6.2.	Next-generation sequencing.....	72
1.7.	Population-based genetic variation datasets: why collect them?.....	73
1.7.1.	Projects for mapping human genetic variation	74
1.7.2.	Clinical uses	77
1.7.3.	Bioinformatics uses	77
1.8.	Summary of Aims and Objectives	78
Chapter 2. Overview of methods.....		80
2.1.	Materials/samples	80
	Participants	80
	Blood samples.....	80
	Buccal swab samples	80
2.2.	Ethics	81
2.3.	Wet-Laboratory methods	82
2.3.1.	DNA extraction and quantification.....	82
	DNA from Blood	82
	DNA from Buccal swabs.....	82
	DNA storage	82
	DNA quality and concentration quantification.....	83
2.3.2.	Polymerase Chain Reaction (PCR)	85
	Primer designing for PCR.....	86
	Using PCR-based methods for variant screening	88
2.3.3.	Gel electrophoresis and visualising PCR products.....	90
	96-well MADGE	90
2.3.4.	Exome targeting	91
2.4.	Bioinformatics and Statistical methods.....	92
2.4.1.	DNA Sequencing and Mapping	92

Whole-exome sequencing	92
Sanger sequencing	92
2.4.2. Variant calling and annotation.....	92
2.4.3. Mutation effect predictors	93
FATHMM.....	94
SIFT	95
PolyPhen-2	95
Condel.....	95
2.4.4. Candidate genes/variants	96
2.4.5. Protein structure modelling.....	96
2.4.6. Calculating <i>F</i> values and identifying autozygous regions	97
2.5. Literature reviews	98
Chapter 3. Identifying highly-penetrant disease-causal mutations from Next-	
generation sequencing data.....	99
3.1. Introduction	100
3.2. Aims and Objectives	102
3.3. Methods	102
3.3.1. Stage 1 - Quality control & Variant calling	102
Targeted sequencing.....	103
Mapping sequence reads	104
Variant calling	106
Additional checks of autozygosity	109
3.3.2. Stage 2 – Filtering/Ranking of Variants	109
Using prior information to rank/filter variants	110
Using effect prediction algorithms to rank/filter variants	111
Further filtering/ranking.....	114
3.3.3. Stage 3 - Building evidence for causality	115
Public data as a source of evidence	118
Mapping causal loci within families	120

Autozygosity mapping	121
Exceptional cases.....	121
Identifying highly penetrant variants for common-complex disorders	122
3.4. Discussion.....	124
3.5. Conclusions	125
Chapter 4. Hunting for Primary ciliary dyskinesia causal genes	126
4.1. Introduction	126
4.1.1. Diagnosis of PCD in Clinical Settings.....	128
4.1.2. Genetic Aetiology of PCD.....	130
4.2. Hypothesis.....	136
4.3. Aims and Objectives	136
4.4. Methods	136
4.4.1. Clinical criteria used for PCD.....	136
Electron Microscopy	137
4.4.2. Participants	137
4.4.3. Collection/storage of samples and DNA extraction	138
4.4.4. DNA sample quality and quantification	138
4.4.5. Whole-exome sequencing.....	139
Quality control on sequencing data	139
4.4.6. Autozygosity mapping	140
4.4.7. Variant prioritisation procedures	140
Candidate genes.....	140
Looking for causal variants	141
4.4.8. Mutation screening and variant validation.....	141
Family 1 and the p.R263* variant in <i>DNALI1</i>	141
Family 2 and the p.R483Q variant in <i>CCT4</i>	142
Family 6, Saudi population and the p.E309* mutation.....	142
4.4.9. Protein structure modelling.....	144

4.5. Results	144
4.5.1. DNA sample quality and concentration.....	144
4.5.2. Family 1	147
WES data statistics	147
Read alignment statistics	148
Variant calling statistics	151
Analysis of whole-exome, LRoH and candidate genes	161
Modelling mutation effect on protein structure	165
Ultrastructure of respiratory cilia	166
4.5.3. Family 2	174
WES data statistics	174
Read alignment statistics	175
Variant calling statistics	177
Analysis of whole-exome, LRoH and candidate genes	192
4.5.4. Family 3	197
WES Data statistics	197
Read alignment statistics	198
Variant calling statistics	199
Analysis of whole-exome, LRoH and candidate genes	204
4.5.5. Family 4	206
WES Data statistics	206
Read alignment statistics	207
Variant calling statistics	208
Analysis of whole-exome, LRoH and candidate genes	213
Modelling mutation effect on protein structure	215
4.5.6. Family 5	216
WES Data statistics	217
Read alignment statistics	217
Variant calling statistics	219
Analysis of whole-exome, LRoH and candidate genes	224
Modelling mutation effect to protein	227

4.5.7. Family 6	228
WES Data statistics	228
Read alignment statistics	229
Variant calling statistics	231
Analysis of whole-exome, LRoH and candidate genes	236
Ultra-structure and Motility of cilia	240
Screening for c.925G>T in Saudi Arabian samples	240
Modelling mutation effect to protein	246
4.6. Discussion.....	247
4.6.1. Whole-exome sequencing.....	247
4.6.2. PCD analyses	249
<i>DNAAF3</i>	250
<i>DNALI1</i>	251
<i>MNS1</i>	252
<i>LRRC48</i>	252
<i>HEATR2</i>	253
<i>CCDC151</i>	253
4.6.3. Anecdotes from literature	255
4.7. Conclusions	256
<i>DNAAF3</i>	256
<i>DNALI1</i>	256
<i>MNS1</i>	257
<i>LRRC48</i> and <i>HEATR2</i>	258
<i>CCDC151</i>	258
Chapter 5. Mutation in <i>ADAT3</i> causes Autosomal recessive Intellectual disability	259
5.1. Introduction	259
5.2. Hypothesis.....	261
5.3. Aims and Objectives	261
5.4. Methods	261
5.4.1. Constructing Family Pedigree	261

5.4.2.	Determining LRoH from SNP array data.....	264
5.4.3.	Brute-force pinpointing of causal region(s)	265
5.4.4.	Haplotype Phasing	266
5.4.5.	Protein structure modelling.....	266
5.5.	Results	266
5.5.1.	<i>F</i> values.....	266
5.5.2.	Autozygosity mapping	267
5.5.3.	Brute force mapping	268
5.5.4.	Haplotype phasing	270
5.5.5.	Modelling mutation effect to protein	271
5.6.	Discussion.....	272
5.6.1.	Information from SNP array data.....	273
5.6.2.	Addition to literature.....	274
5.7.	Conclusions	275
Chapter 6. Proxy molecular diagnosis from whole-exome data reveals Papillon-Lèfevre syndrome causal mutation		
		276
6.1.	Introduction	276
6.2.	Hypothesis.....	277
6.3.	Aims and Objectives	277
6.4.	Methods	277
6.4.1.	Ethics.....	277
6.4.2.	Probability of proxy diagnosis	277
6.4.3.	Participants and Genetic Data Analysis	278
6.4.4.	PCR amplification and Sanger sequencing	279
6.4.5.	Mutation screening using ARMS-PCR	279
	Local Population	279

Family members.....	279
6.4.6. Protein structure modelling.....	280
6.5. Results.....	280
6.5.1. Whole-exome sequencing of PCD affected sibling	280
6.5.2. Screening <i>CTSC</i> gene.....	280
6.5.3. Cathepsin C protein structure prediction.....	286
6.5.4. Screening for p.G300D variant in Riyadh, KSA	287
6.6. Discussion.....	292
6.6.1. On therapy and cures	293
6.6.2. Addition to literature.....	293
6.6.3. Considerations.....	294
6.7. Conclusions	294
Chapter 7. Importance of consanguineous populations to characterization of human gene function	295
7.1. Introduction	295
7.2. Hypothesis.....	300
7.3. Aims and Objectives	300
7.4. Natural human gene knock-outs in consanguineous populations	301
7.4.1. Overview of literature	304
7.4.2. Effects of consanguinity on Mendelian disease.....	306
7.4.3. Effects of consanguinity on common-complex diseases	308
7.5. <i>Quasi</i> reverse genetics studies in humans?.....	308
7.5.1. Frequency of natural gene knock-outs.....	309
7.5.2. Suitable Sample Populations.....	318
7.6. Methods	322
7.6.1. Ethics.....	322

7.6.2.	Creating a DNA Bank.....	323
7.6.3.	Identification of Autozygous regions and Gene inactivating variants	323
7.6.4.	Comparative Genomics.....	324
7.7.	Discussion.....	324
7.7.1.	Way forward?	326
7.7.2.	Addition to literature.....	327
7.7.3.	Considerations.....	329
7.8.	Conclusions	329
Chapter 8. Overall Discussion and Conclusions		330
8.1.	Discussion on disease causal genes and WES.....	331
8.2.	Considerations and Thoughts on Consanguinity	334
8.3.	General additions to literature	337
8.4.	Overall Conclusions and Future work.....	338
Chapter 9. Appendices		339
9.1.	General appendices	339
9.2.	Baha'ism's view on consanguinity	341
9.3.	Milestones in Genetics research	342
9.4.	Appendices for Chapter 3 – NGS data analysis guide.....	344
9.5.	Appendices for Chapters 4 and 6 – Families with PCD	347
9.5.1.	PCD sample quality	347
9.5.2.	All known and potential PCD causal genes.....	348
9.5.3.	Looking for causal variants	360
9.5.4.	Files used in PCD analyses	361
9.5.5.	Scripts used in PCD analyses	367
9.5.6.	Protein structure modelling.....	368

PDB data format.....	368
<i>MNS1</i>	368
<i>DNALI1</i>	374
<i>DNAAF3</i>	379
<i>CCDC151</i>	384
<i>CTSC</i>	389
9.6. Appendices for Chapter 5 - Intellectual disability analyses.....	394
9.6.1. Protein structure modelling.....	394
<i>ADAT3</i>	394
9.6.2. Autozygosity mapping within family	399
9.6.3. ARID causal/associated genes from GeneCards	403
9.6.4. Details on all ARID causal genes/mutations from literature.....	408
9.7. Consanguineous Unions	430
9.8. Other scripts, Files and UNIX commands used.....	434
Bibliography	448

Table of Tables

Table 1.1	Views of main religions towards consanguineous marriages	49
Table 1.2	Clinical potential of widely used methods	70
Table 2.1	Nanodrop results for randomly selected 9 samples	83
Table 2.2	Qubit Fluorometer results for the same 9 samples	83
Table 3.1	Tools for aligning reads to a reference genome	105
Table 3.2	Tools for identifying variation from a reference genome using NGS reads	108
Table 3.3	Tools for predicting variant effects: Identifying neutral and pathogenic mutations	113
Table 3.4	What is needed for a genetic study?	125
Table 4.1	Currently known human PCD causal/associated genes and/or regions	133
Table 4.2	Primers used to amplify a 235bp long region containing the p.R263* mutation in the <i>DNALI1</i> gene	141
Table 4.3	Primers used to amplify a 238bp long region containing the p.R483Q mutation in the <i>CCT4</i> gene	142
Table 4.4	Primers used to amplify a 221bp long region containing the p.E309* mutation in the <i>CCDC151</i> gene	143
Table 4.5	Digestion of <i>CCDC151</i> amplicons using <i>AvrII</i> enzyme	143
Table 4.6	DNA sample test results for 9 individuals	145
Table 4.7	Family 1's WES data quality summary statistics	148
Table 4.8	Family 1's read alignment quality summary statistics	149
Table 4.9	Family 1 Individual 1's SNP summary statistics	151
Table 4.10	Family 1 Individual 1's indel summary statistics	151
Table 4.11	Family 1 Individual 2's SNP summary statistics	155
Table 4.12	Family 1 Individual 2's indel summary statistics	156
Table 4.13	Local sequence alignment containing the mutated residue from multiple alignment of the <i>DNALI1</i> gene in different organisms	166
Table 4.14	Family 2's WES data quality summary statistics	174
Table 4.15	Family 2's alignment quality summary statistics	175
Table 4.16	Family 2 Individual 1's SNP summary statistics	176
Table 4.17	Family 2 Individual 1's indel summary statistics	177
Table 4.18	Family 2 Individual 2's SNP summary statistics	181
Table 4.19	Family 2 Individual 2's indel summary statistics	182
Table 4.20	Family 2 Individual 3's SNP summary statistics	186

Table 4.21	Family 2 Individual 3's indel summary statistics	187
Table 4.22	Family 3 Individual 1's WES data quality summary statistics	196
Table 4.23	Family 3 Individual 1's alignment quality summary statistics	197
Table 4.24	Family 3 Individual 1's SNP summary statistics	198
Table 4.25	Family 3 Individual 1's indel summary statistics	199
Table 4.26	Family 4 Individual 1's WES data quality summary statistics	205
Table 4.27	Family 4 Individual 1's alignment quality summary statistics	206
Table 4.28	Family 4 Individual 1's SNP summary statistics	207
Table 4.29	Family 4 Individual 1's indel summary statistics	208
Table 4.30	Family 5 Individual 1's WES data quality summary statistics	216
Table 4.31	Family 5 Individual 1's alignment quality summary statistics	217
Table 4.32	Family 5 Individual 1's SNP summary statistics	218
Table 4.33	Family 5 Individual 1's indel summary statistics	219
Table 4.34	Family 6 Individual 1's WES data quality summary statistics	228
Table 4.35	Family 6 Individual 1's alignment quality summary statistics	229
Table 4.36	Family 6 Individual 1's SNP summary statistics	230
Table 4.37	Family 6 Individual 1's indel summary statistics	231
Table 4.38	Local sequence alignment containing the mutated residue from multiple alignment of the <i>CCDC151</i> gene in different organisms	238
Table 5.1	SNPs tagging the ARID causal region	267
Table 6.1	Local sequence alignment containing the mutated residue from multiple alignment of the <i>CTSC</i> gene in different species	282
Table 6.2	Primers used in ARMS-PCR for genotyping the p.G300D variant	287
Table 7.1	A comparison between collections of outbred offspring and collections of offspring of first cousins	312
Table 7.2	A comparison between collections of outbred offspring and collections of offspring of uncle-niece unions	313
Table 7.3	Potential number of LoF mutations in the human genome	316
Table 10.1	All known human PCD causal genes and the variants which cause them	Ap.
Table 10.2	Ensembl IDs of all genes found in the Ciliome database	Ap.
Table 10.3	Legend for the 10 columns of data in the PDB file	Ap.
Table 10.4	Domains used for Ginzu Prediction when predicting MNS1p structure	Ap.
Table 10.5	Domains used for Ginzu Prediction when predicting DNALI1p structure	Ap.

Table 10.6	Domains used for Ginzu Prediction when predicting DNAAF3p structure	Ap.
Table 10.7	Domains used for Ginzu Prediction when predicting CCDC151p structure	Ap.
Table 10.8	Domains used for Ginzu Prediction when predicting CTSCp structure	Ap.
Table 10.9	Domains used for Ginzu Prediction when predicting ADAT3p	Ap.
Table 10.10	LROHs identified from each affected individual using David Pike's method	Ap.
Table 10.11	Examples of known ARID causal genes	Ap.
Table 10.12	All known human ARID causal genes and the variants which cause them	Ap.

Table of Figures

Figure 1.1	Anatomy of the human cell	2
Figure 1.2	Structure of DNA	3
Figure 1.3	Central Dogma of Molecular Biology	7
Figure 1.4	Types of single nucleotide variants	11
Figure 1.5	Types of indels	14
Figure 1.6	Sequence ontology terms from the Ensembl Variant Effect Predictor	19
Figure 1.7	Autosomal recessive pattern of inheritance	22
Figure 1.8	Contrast between recessive and dominant mutations	24
Figure 1.9	Relationship between allele frequencies, penetrance and discovery	29
Figure 1.10	Factors influenced by consanguinity and culture	44
Figure 1.11	Homozygosity of two identical by descent alleles	51
Figure 1.12	Example of a complex pedigree with multiple consanguineous unions	54
Figure 1.13	Worldwide consanguinity	57
Figure 1.14	Laws regarding first-cousin marriage around the world	59
Figure 2.1	Schematic drawing of the PCR cycle	85
Figure 2.2	ARMS-PCR methodology explained	89
Figure 3.1	Steps in whole-exome sequencing	103
Figure 3.2	Post-VCF file procedures	110
Figure 3.3	Finding ‘the lot’ in Complex disorders: Searching for causal variants	116
Figure 3.4	Finding ‘the one’ in Mendelian Disorders: Searching for the causal variant	116
Figure 3.5	Filtering steps applied to all mutations in the exome	119
Figure 3.6	Summary of whole analysis process: DNA sample to identification of variant	123
Figure 4.1	Cross-sectional depiction of respiratory cilia	128
Figure 4.2	Workflow for diagnosing PCD as suggested by Busquets <i>et al</i>	129
Figure 4.3	Electrophoretogram result for DNA sample integrity	146
Figure 4.4	Family 1 Individual 1’s indel length distribution	153
Figure 4.5	Family 1 Individual 1 CNV distribution	154
Figure 4.6	Family 1 Individual 2’s indel length distribution	158
Figure 4.7	Family 1 Individual 2’s CNV distribution	159

Figure 4.8	Plotting of variant status across the genome for individual 1	162
Figure 4.9	Plotting of variant status across the genome for individual 2	163
Figure 4.10	Protein structure of wild type DNALI1 protein	164
Figure 4.11	Protein structure of mutant DNALI1 protein	165
Figure 4.12	Proteins DNALI1p is predicted to interact with	167
Figure 4.13	Reads mapped to the reference human genome at the site of the p.R263* mutation	168
Figures 4.14A-C	EM image of (A) Control cilia (B and C) individual 1's cilia	169
Figures 4.14D-E	EM image of (A) Control cilia (D and E) individual 2's cilia	170
Figure 4.15	Confirmation of variant status in proband and other family members using Sanger sequencing	171
Figure 4.16	Filtering steps applied to all mutations in the exome of proband	172
Figure 4.17	Family 2 Individual 1's indel length distribution	179
Figure 4.18	Family 2 Individual 1 CNV distribution	180
Figure 4.19	Family 2 Individual 2's indel length distribution	184
Figure 4.20	Family 2 Individual 2 CNV distribution	185
Figure 4.21	Family 2 Individual 3's indel length distribution	189
Figure 4.22	Family 2 Individual 3 CNV distribution	190
Figure 4.23	Plotting of variant status across the genome for Family 2 Individual 1	193
Figure 4.24	Plotting of variant status across the genome for Family 2 Individual 2	194
Figure 4.25	Plotting of variant status across the genome for Family 2 Individual 3	195
Figure 4.26	Family 3 Individual 1's indel length distribution	201
Figure 4.27	Family 3 Individual 1's CNV distribution	202
Figure 4.28	Plotting of variant status across the genome for Family 3 Individual 1	204
Figure 4.29	Family 4 Individual 1's indel length distribution	210
Figure 4.30	Family 4 Individual 1 CNV distribution	211
Figure 4.31	Plotting of variant status across the genome for Family 4 Individual 1	213
Figure 4.32	Protein structure of wild type DNAAF3 protein	214
Figure 4.33	Protein structure of mutant DNAAF3 protein	215
Figure 4.34	Family 5 Individual 1's indel length distribution	221
Figure 4.35	Family 5 Individual 1 CNV distribution	222

Figure 4.36	Plotting of variant status across the genome for Family 5 Individual 1	224
Figure 4.37	Reads mapped to the reference human genome at the site of the p.M263T mutation	225
Figure 4.38	Protein structure of wild type MNS1 protein	226
Figure 4.39	Protein structure of mutant MNS1 protein	227
Figure 4.40	Family 6 Individual 1's indel length distribution	233
Figure 4.41	Family 6 Individual 1 CNV distribution	234
Figure 4.42	Plotting of variant status across the genome for Family 6 Individual 1	236
Figure 4.43	Filtering steps applied to all mutations in the exome	237
Figures 4.44A-B	Cross-sections of respiratory cilia in (A) control and (B) CCDC151 mutated proband	240
Figure 4.45	Confirmation of variant status in proband and other family members using Sanger sequencing and <i>AvrII</i> digestion	241
Figures 4.46A-C	Screening the Saudi population for the p.E309* variant	243
Figure 4.47	Reads mapped to the reference human genome at loci of the c.924C>A (p.E309*) mutation	244
Figure 4.48	Protein structure of wild type CCDC151 protein	245
Figure 4.49	Protein structure of mutant CCDC151 protein	246
Figure 5.1	Whole family pedigree of participating family	262
Figure 5.2	Using autozygosity mapping to pinpoint ARID causal loci	268
Figure 5.3	Setting LRoH boundaries between the affected individuals in the two families	269
Figure 5.4	Protein structure of wild type ADAT3 protein	270
Figure 5.5	Protein structure of mutant ADAT3 protein	271
Figure 6.1	Reads mapped to the human reference genome at the p.G300D mutation loci	281
Figure 6.2	Confirmation of variant status in other family members using Sanger sequencing	283
Figure 6.3	Validation of p.G300D in all family members using ARMS-PCR	284
Figure 6.4	Protein structure of wild type CTSC protein	285
Figure 6.5	Protein structure of mutant CTSC protein	286
Figures 6.6A-F	Screening the local population for the c.899G>A:p.(G300D) variant	288
Figure 7.1	Examples of inferences to be gained from autozygous regions in consanguineous offspring	290
Figure 7.2	Example of differences between union of unrelated and consanguineous individuals	302
Figure 7.3	Consanguinity and increased homozygosity due to recent common ancestor	306

Figure 7.4	Comparison between offspring of outbred individuals and first cousins using the example of an allele for which $q = 0.1$	314
Figure 7.5	Comparison between offspring of outbred individuals and first cousins using the example of an allele for which $q = 0.001$	315
Figure 7.6	Location of Riyadh in the Arabian Peninsula and KSA	319
Figure 7.7	Locations of Andhra Pradesh and Karnataka in India	320
Figure 10.1	The Genetic Code	Ap.
Figure 10.2	Certificate confirming HTA training received	Ap.
Figure 10.3	Letter received from the National Spiritual Assembly of the Bahá'ís of the United Kingdom detailing Baha'ism's view on consanguinity	Ap.
Figure 10.4	PCD Sample 5 tested for integrity	Ap.

List of Equations

Equation 2.1	Calculating the inbreeding coefficient of an individual (F_o)	96
Equation 2.2	Simpler version of Wright's inbreeding coefficient formula	97
Equation 7.1	The Hardy-Weinberg equation (HWE)	296
Equation 7.2	Calculating total number gene inactivations in a consanguineous collection	310

List of peer-reviewed publications

The below are the journal articles and conference papers/abstracts published as a result of this thesis (chronological order):

- 1- Erzurumluoglu *et al.*, Oct 2014. Nonsense mutation in *CCDC151* causes Primary ciliary dyskinesia. Molecular Basis of Mendelian Disorders, ID: 2932S. American Society of Human Genetics (ASHG) conference 2014. San Diego, CA, USA.
- 2- Alsaadi and Erzurumluoglu *et al.*, Dec 2014. Nonsense mutation in coiled coil domain containing 151 gene (*CCDC151*) causes Primary ciliary dyskinesia. Human Mutation. 35 (12). doi: 10.1002/humu.22698
- 3- Erzurumluoglu *et al.*, Mar 2015. Identifying highly-penetrant disease causal mutations using next generation sequencing: Guide to whole process. BioMed Research International. 2015 (2015). doi: 10.1155/2015/923491
- 4- Erzurumluoglu *et al.*, Mar 2015. Proxy molecular diagnosis from whole-exome sequencing reveals Papillon-Lèfevre syndrome caused by a missense mutation in *CTSC*. PLoS One. 10 (3). doi: 10.1371/journal.pone.0121351
- 5- Erzurumluoglu *et al.*, Dec 2015. Importance of genetic studies in consanguineous populations to characterization of human gene function. Annals of Human Genetics. Accepted: 21/12/15.

Author Contributions

- 1- AME produced poster and content with guidance from SR, TRG and INMD (Conference abstract and poster).
- 2- AME wrote the manuscript (with guidance from SR, TRG and INMD). AME carried out *in silico* and wet-lab analyses. INMD and MMA led the study; and together with SR, KKA, PAIG and TRG, provided guidance throughout study and also commented on the manuscript. MMA carried out diagnosis and obtained consent from family. ACA, MM, HZO and MMA led the collection and processing of EM images for cilia. PAIG and AME performed DNA

extraction, quantification and other DNA quality control procedures. All authors approved final version of manuscript.

- 3- AME wrote the manuscript (with guidance from SR, TRG and INMD). TRG led the study; and SR, INMD, HAS, TGR and DB provided advice and comments. All authors approved final version of manuscript.
- 4- AME wrote the manuscript (with guidance from SR, TRG and INMD). INMD led the study; and together with SR, KKA, MMA, PAIG and TRG, provided guidance throughout study and also commented on the manuscript. AME, SL, AG and TSA carried out wet-lab analyses. FMA, MMA and BMA provided clinical diagnosis and obtained consent from family. All authors approved final version of manuscript.
- 5- AME wrote the manuscript (with guidance from SR, TRG and INMD). INMD led the study; and together with SR, HAS and TRG, provided guidance throughout study and also commented on the manuscript. All authors approved final version of manuscript.

Signatures (of all first and corresponding authors):

A. Mesut Erzurumluoğlu Ian N.M. Day Tom R. Gaunt Muslim M. Alsaadi

List of other peer-reviewed publications

The below are the journal articles and conference papers published during the course of this PhD, not related to this thesis (chronological order):

- 1- Erzurumluoglu *et al.*, 2015. Novel uses of Y chromosomal haplogroups in genetic association studies and suggested implications. Submitted.

List of non-academic publications

The below are the non-academic articles published as a result of this thesis (chronological order, all in English unless stated otherwise):

- 1- *Erzurumluoglu, M. July 2012. Consanguineous Marriages in the Light of New Technological Developments. Hiyerarşi. Source URL:
<https://mesuturkey.wordpress.com/2012/07/01/consanguineous-marriages-in-the-light-of-new-technological-advancements/>
- 2- Erzurumluoglu, M. May 2014. Consanguineous Marriages: Perspectives from Social Taboos, Religion, and Science. The Fountain. Issue #99. Source URL:
<http://www.fountainmagazine.com/Issue/detail/consanguineous-may-2014>
- 3- †Erzurumluoglu, A.M. Jan 2015. Akraba evlilikleriyle ilgili objektif perspektif ve ilginç anekdotlar. Blog URL:
<https://mesuturkey.wordpress.com/2015/01/29/akraba-evlilikleriyle-ilgili-objektif-perspektif-ve-ilginc-anekdotlar/>
- 4- Erzurumluoglu, A.M. Sept 2015. What connects history, culture, law, genetics and preventive medicine? Blog URL:
<https://mesuturkey.wordpress.com/2015/09/22/what-connects-history-culture-law-genetics-and-preventive-medicine/>

*English and Turkish versions available

†Turkish

Table of commonly used acronyms and/or abbreviations

Acronym	Definition
1000GP	1000 Genomes Project
Φ mutations	Mutations with (predicted) ‘high deleterious impact’ *At the protein level
ARID	Autosomal recessive Intellectual disability *Also known as Mental retardation in the literature
BAM	Binary Sequence Alignment/Map format
BLAST	Basic Local Alignment Search Tool *Algorithm for comparing query sequences with sequences from other species
BWA	Burrows-Wheeler Aligner
CNV	Copy number variation
dbSNP	NCBI SNP database
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide triphosphate *Generic term for the 4 deoxyribonucleotides: dATP, dCTP, dGTP, dTTP
EM	Electron Microscope
EVS	Exome Variant Server *Also known as NHLBI GO Exome Sequencing Project
ExAC	Exome Aggregation Consortium *Contains MAF data from 60,706 individuals
FATHMM	Functional Analysis Through Hidden Markov Models *A mutation effect predictor which makes use of Hidden Markov Models
F value	Wright’s inbreeding coefficient
GERP	Genomic Evolutionary Rate Profiling *A score indicating how conserved a nucleotide/amino acid is
GWAS	Genome-wide association study
HGP	Human Genome Project
IBD	Identical by descent *Allele(s) inherited from a recent common ancestor
IDA/ODA	Inner dynein arm/Outer dynein arm
Indel	Small insertion or a deletion event in a genome *Umbrella term for both
LD	Linkage Disequilibrium
LoF	Loss of Function
LRoH	Long Run of Homozygosity *Used mostly as a proxy for autozygous region in consanguineous offspring
KSA	Kingdom of Saudi Arabia
MAF	Minor Allele Frequency *Will mostly use global MAF from 1000GP
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NMD	Nonsense mediated decay
OMIM	Online Mendelian Inheritance in Man
PCD	Primary ciliary dyskinesia
PCR	Polymerase Chain Reaction

PLS	Papillon-Lèfevre syndrome
PolyPhen	Polymorphism Phenotyping *One of the most commonly used mutation effect predictors
RNA	Ribonucleic Acid
QRG	<i>Quasi</i> Reverse Genetics (studies in human populations) *Name given to proposed study in consanguineous populations
SAM	Sequence Alignment/Map format
SIFT	Sorting Intolerant From Tolerant *A commonly used mutation effect predictor
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant *SNVs with MAF $\geq 1\%$ are SNPs
sSNV / nsSNV	Synonymous SNV / Nonsynonymous SNV *SNVs which do not cause an amino acid change in the primary structure of a protein / SNVs which do
VCF	Variant Call Format
VEP	Ensembl Variant effect predictor *Commonly used variant consequence annotation tool
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

CHAPTER 1. INTRODUCTION AND LITERATURE REVIEW

In the beginning there was nothing*, and then came the first cell. Therefore I shall also start this thesis with the cell. I follow on by giving an overview of the human genome and the biological processes linking the genome to (Mendelian and complex) disease phenotypes. I introduce important concepts such as mutation, transcription and translation in section 1.1 before moving onto introducing the area of Genetic Epidemiology in section 1.2, which is the study of genetic causes and effects in specified human populations. I then introduce the main concept of this thesis in section 1.3 – consanguinity, and how it relates to human disease genetics research. Furthermore, in sections 1.4, 1.5, 1.6 and 1.7, I briefly present how genetic analysis is carried out with the use of next generation sequencing (NGS) platforms. I also provide a historical and/or traditional background to the topics where appropriate. Finally, I end the chapter with the overall objectives of this thesis.

1.1. Organism to Genome to Genes

All ‘living’ things (organisms hereafter) on Earth are made up of cells with numbers ranging from a single (e.g. bacteria) to trillions (e.g. humans). Therefore cells are called the ‘building blocks’ of organisms; and humans are no different from other organisms in this respect with over 10 trillion (10^{13}) cells [1, 2] working side by side for the correct functioning of the whole system, ultimately enabling life to carry on. There is a vast array of the types of human cells with sizes ranging from $10\mu\text{m}$ (10^{-6} metres) to over $100\mu\text{m}$ (10^{-5}m) in size [2]; and most cell types cannot be seen with the naked eye. However, in contrast to their microscopic sizes, every single human cell could be seen as ‘a universe on its own’ with many organelles residing within its

* No living things on Earth

outer membrane (Figure 1.1). These organelles carry out specific jobs, and are needed to interact with each other at the right place and time. Remove one and the cell would be a chaos.

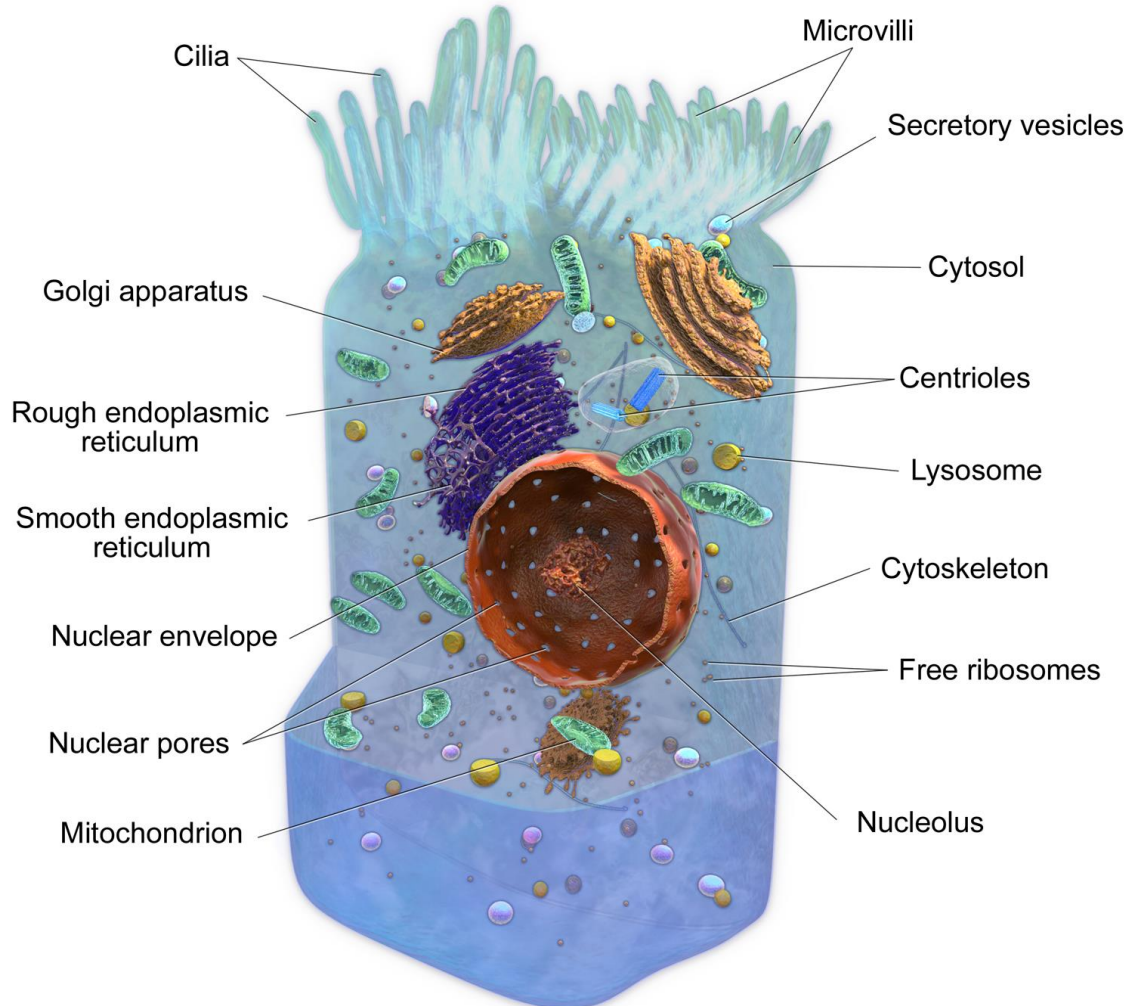


Figure 1.1 Anatomy of a typical human cell. Mitochondria (plural form of Mitochondrion) are the only organelles to have their own DNA outside of the nucleus, although they still require around a 1000 nuclear-encoded proteins to function properly [3-5]. Image reproduced under the Creative Commons Licence, source URL: [http://en.wikipedia.org/wiki/Cell_\(biology\)](http://en.wikipedia.org/wiki/Cell_(biology))

At the core of each human cell there is a compartment called the ‘nucleus’, which etymologically derives from the Latin for ‘kernel’ [6]. The nucleus is home to the ‘instructions’ (i.e. genetic blueprint) on how to build an entire human being. These instructions are in the form of deoxyribonucleic acid (DNA) which is virtually found in all life on Earth (Figure 1.2). In most species which reproduce sexually, humans being one of them, DNA is passed on from both parents to the offspring enabling the

continuation of the species. It is a double stranded, helical shaped polymer of nucleotides which itself is composed of a ribose sugar, phosphate group and one of the four nitrogenous bases, adenine (abbreviation: A), cytosine (C), guanine (G) and thymine (T) [7, 8]. The human genome is diploid and is made up of a total of 46 chromosomes with 22 pairs of autosomes (named chromosomes 1 to 22) and a pair of sex (determining) chromosomes (named X and Y) [9]. Normally, the offspring either have two copies of the X chromosome or a copy each of X and Y (from mother and father respectively), causing the offspring to be a female or a male respectively.

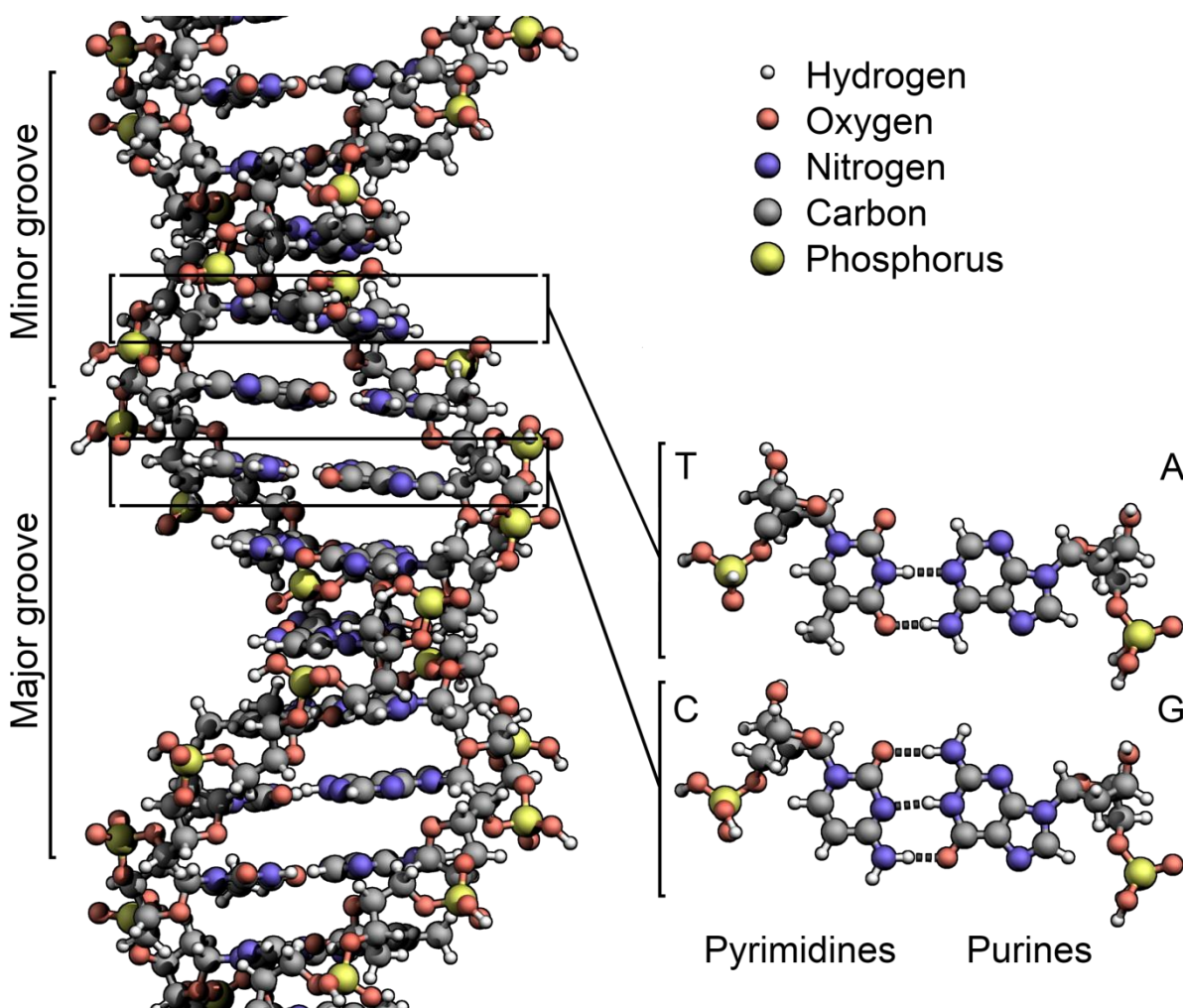


Figure 1.2 Structure of Deoxyribonucleic Acid (DNA). The human genome is approximately 3.3 billion base pairs long. Image reproduced under the Creative Commons License, source URL: <http://en.wikipedia.org/wiki/DNA>

Mitochondria, one of the essential organelles in the cell (i.e. the “power plant” of the cell [10]), also has its own genome thus is also included in many definitions of the

‘human genome’. The haploid human genome is approximately 3.3 billion (3.3×10^9) base pairs long (according to Ensembl Assembly: GRCh37.p12, Feb 2009 available at: http://www.ensembl.org/Homo_sapiens/Info/Annotation#assembly), however it is no bigger than a speck of dust and it is estimated that a single gram of DNA will store around two petabytes (2×10^{15} bytes) of data [11].

1.1.1. What is a gene?

Historical background

Ancient Greek philosophers (especially Aristotle and Hippocrates) and many great thinkers from later civilisations such as the early Islamic scholars Al-Jahiz* and Abu al-Qasim al-Zahrawi† (also known as Albucasis), and Sephardic physician (also poet and philosopher) of the 12th century Jusah HaLevi‡ acknowledged that offspring resembled their parents and certain disease traits were more heritable than others; and they speculated about possible mechanisms of heredity (e.g. semen, blood, diet, materials from around the body, combinations) [12-14].

This was all we knew until a set of revolutionary experiments were carried out by Gregory Mendel. He arrived on to the scene with his pea-plant experiments and conclusions in 1866. These conclusions were later named ‘the three basic laws of heredity’ (i.e. laws of independent segregation, independent assortment and dominance respectively) [15]. Six years later, Friedrich Miescher extracted ‘nucleins’ (nucleic acids) from white blood cells, which paved the way for DNA to be recognised as the carrier of inheritance [16]. Until 1910 there were fierce debates about whether ‘the cause of heritable phenotypes’ (i.e. genes) resided on chromosomes or not. The disputes were settled by Thomas Morgan’s findings from his fruit fly (*Drosophila melanogaster*) studies demonstrating the existence of sex chromosomes and sex-linked traits [17]. As the chromosome is made up of proteins

* Studied the effect of the environment on the survival rates of animals

† First physician to identify the hereditary nature of haemophilia

‡ Described some dominant and recessive traits

as well as DNA, many thought that the former was the hereditary material; and 1944 was the year when DNA (and not protein) was proven to be hereditary material using genetic transformation in bacteria [18]. Then came many other discoveries (see section 9.3) and consequent challenges including the task of understanding how the molecular information in the genome of an individual translated into phenotypes. These challenges were mainly solved by the discovery of the DNA transcription and translation mechanisms – and the whole process of translating DNA to a protein (detailed below) was baptised ‘the central dogma of molecular biology’ by Francis Crick [19].

The modern ‘gene’

Arguably, the most eye-catching project of the century was the Human Genome Project (HGP) which started in 1993 and lasted ten years, helping us to decode and understand our own genome. The HGP helped us identify how many ‘genes’ there were – regions in the genome which ultimately code for proteins through its sequence (of bases, A, C, G or T) being transcribed to RNA (messenger RNA, mRNA) via the transcription process, and then translated to amino acids (depending on the triplet of nucleotides, called codons), and ultimately to proteins, via the mechanism of ‘translation’ through the use of transfer RNA (tRNA, Figure 1.3). This is why genes are called ‘coding’ regions of the genome; and DNA sequences outside of genes are called ‘noncoding’, although there are also noncoding regions within genes called introns. This is not to say noncoding DNA is ‘junk’ as was once believed, as there is increasing evidence of the importance of noncoding DNA such as in gene expression regulation and alternative splicing [20]. As noncoding DNA is not within the scope of this thesis, it will be discussed no more (for comprehensive reviews on noncoding DNA and its importance to human health, see references [21, 22]).

Genes in the human (and virtually all other organisms’) genome start with an ATG triplet (i.e. start codon, a methionine residue) and the order of triplets of bases subsequent to the start codon determines which amino acids will be coded for – until

the sequence is stopped by either one of the TAG, TAA or TGA triplets (i.e. stop codons). Using this definition combined with experimental data, it is estimated that there are around 21000 genes in the human genome [23]. Variation in the DNA sequence of these genes is the major cause of the (heritable) phenotypic differences observed between individuals (and between populations).

The ENCODE project has provided an update to the definition of a gene as 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products', thus non-coding ribonucleic acids (ncRNA) and (non-genic) highly conserved regions of the genome are also included as functional units and therefore, potential genes [24-26].

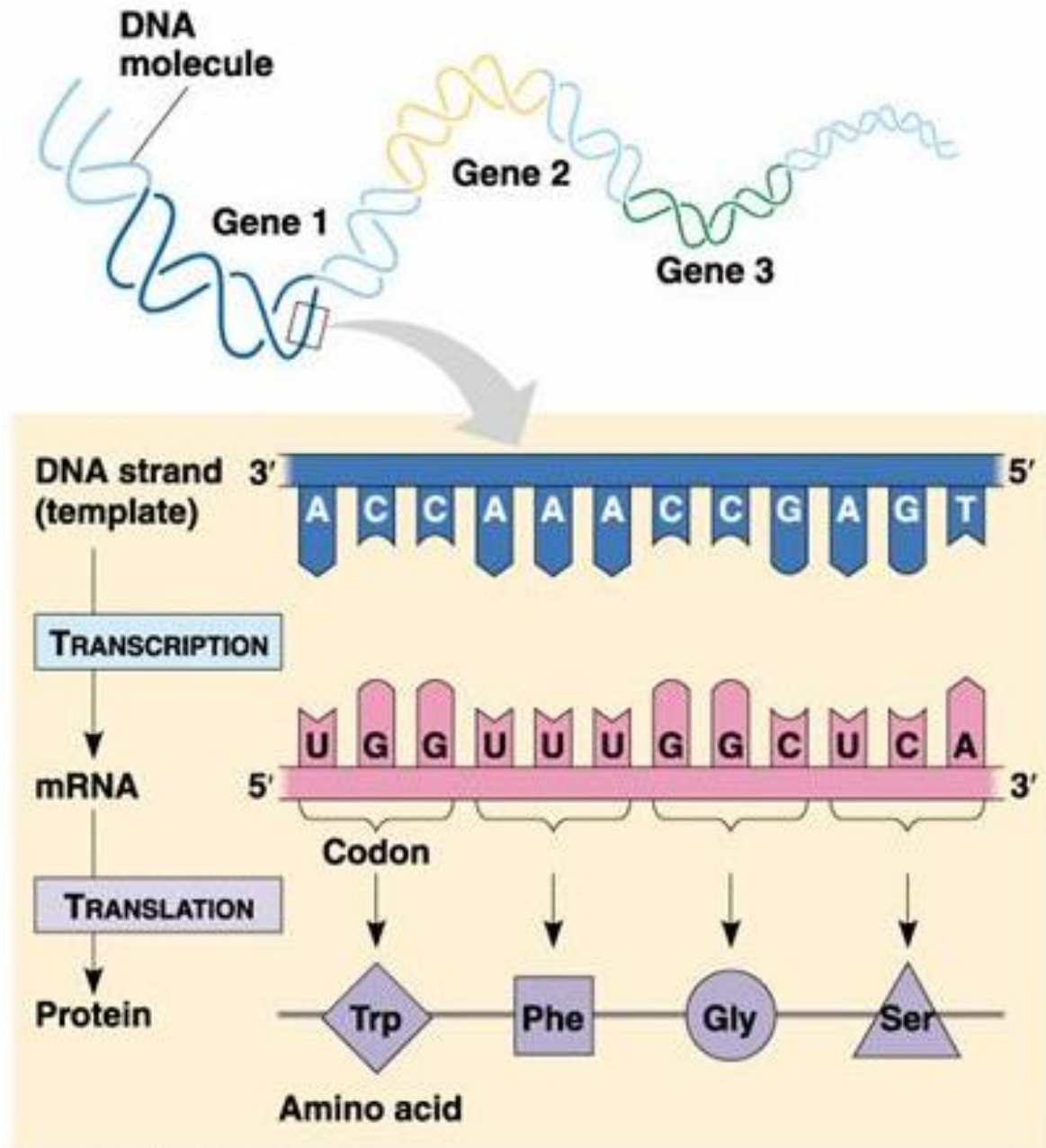


Figure 1.3 The Central Dogma* of Molecular Biology: DNA to RNA to Protein through the mechanisms of transcription and translation respectively. Which codon (i.e. triplet of DNA bases) corresponds to which amino acid is listed in the genetic code table (see Figure 10.1). This image was adapted under the Creative Commons License, source URL: <http://i308.wikispaces.com/>

* The term 'dogma' does not reflect the reality of molecular biology completely. For example reverse transcription (RNA to DNA) does not fit the model. Also post-transcriptional (e.g. alternative splicing, RNAi) and post-translational (e.g. acetylation, phosphorylation) modifications are not accounted for in this 'traditional' model. However these exceptions will not be discussed further as they are not within the scope of this thesis.

1.1.2. Variation in a genome

There is a great deal of variation* in each individual's genome relative to any other individual's, even when within the same family [27, 28]. The only exceptions to this rule are identical (i.e. monozygotic) twins; and even they have *de novo* mutations which separates them genetically† (mutation rate: $\sim 1.1 \times 10^{-8}$ per position per haploid genome [29, 30]). This phenomenon of huge variation is attributable to the human genome's enormous size, abundance of environmental mutagens, and occasional mistakes in the DNA replication and DNA repair processes. However a large majority of these variations have no clinical importance (i.e. harmless) and/or have minor/subtle effects such as changing eye/hair/skin colour or influencing weight/height of an individual; hence we owe virtually all heritable diversity amongst human populations (e.g. ethnicities, uniqueness of each individual) we observe to these harmless variants. Some can even be beneficial (e.g. provide resistance against certain pathogens [31]). Other side of the coin however is that even a single one of these variants which fall in a 'critical' region may be one too many as it can be detrimental to the individual harbouring the mutation (such as in references [32-35]). Unearthing the genetic component of any disease can lead to a better understanding of what went wrong, the mechanisms and the (other) genes involved; and may ultimately lead to preventive interventions and/or curative therapies subsequently.

With the completion of the HGP, where the entire genomes of multiple (healthy) individuals was sequenced, a 'reference human genome' was established [36]. The reference genome is a consensus of the sequences of these (anonymous) individuals and it has enabled researchers to compare their query DNA sequences (e.g. obtained from individuals affected with a particular disease) with it and analyse the differences as 'causal' candidates. The HGP cost a staggering \$2.7 billion of public

* Although approx. 99.9% of two unrelated individuals are identical at nucleotide level, considering the nine figure size of the human genome, the remaining 0.1% is still a large number ($0.001 \times 3.3 \times 10^9 = 3.3$ million).

† $(1.1 \times 10^{-8}) \times (3.3 \times 10^9) \times 2 \approx 70$ *de novo* mutations per diploid genome per generation

money and 13 years' time of many labs around the world to complete; but now the same procedure can be done with less than \$10000 [36-39]. With the ever decreasing prices of sequencing and the advances in computing and technology, it is now possible for researchers to identify the genetic basis of many diseases and health related traits in a very short time (less than a month, sometimes in a day*) [40, 41].

Heritable variants

A variant, in its broad sense, can be classified as **any** difference with respect to the reference sequence; thus it can have different forms and sizes. For example it can affect a single nucleotide (e.g. SNPs) or a large region of a chromosome (e.g. chromosomal aberrations); and the nature of its effects will depend on the type of change (e.g. consequence of base change), where (e.g. which gene? where in the gene? in a somatic cell or a germ cell?) and when (e.g. during gametogenesis or not?) it has occurred. A single nucleotide change in a coding region can result in a synonymous or a nonsynonymous change (Figure 1.4) [42, 43]. Definitions and examples are as follows:

- 1- Synonymous SNVs (sSNVs) are SNVs which bring about a change in the DNA and mRNA sequence of a gene, but do not affect the amino acid sequence of a protein. For example, substituting TTA with TTG will not affect the coding amino acid residue. TTA and TTG both code for Leucine. Synonymous mutations are a consequence of the codon degeneracy[†] in the genetic code.
- 2- Missense SNVs alter the DNA and mRNA sequence of a gene, and also the amino acid sequence of a protein. For example, changing the final thymine

* To read about the 'Milestones in genetics research' which paved the way for today's technologies and findings, please see section 9.3 in the appendices chapter

[†] Multiple codons encoding the same amino acid residue. Although there are 64 different combinations of triplets of the four bases (4³), only 20 different amino acids are encoded.

base in TTT to an adenine base (i.e. to TTA), will cause a change from Phenylalanine to a Leucine in the amino acid sequence.

- 3- Nonsense SNVs change the nucleotide sequence of a gene and mRNA such that the existing amino acid residue now encodes for a stop codon. This ultimately causes a truncated protein to be synthesised – if at all*.

* If the stop gain (i.e. nonsense mutation) occurs towards the 5' side of a gene (or N terminus side of the predicted protein), where more than 50bp remains on the penultimate exon after the mutation, the transcript may be a target for nonsense mediated decay – hence causing the protein to be not synthesised at all.

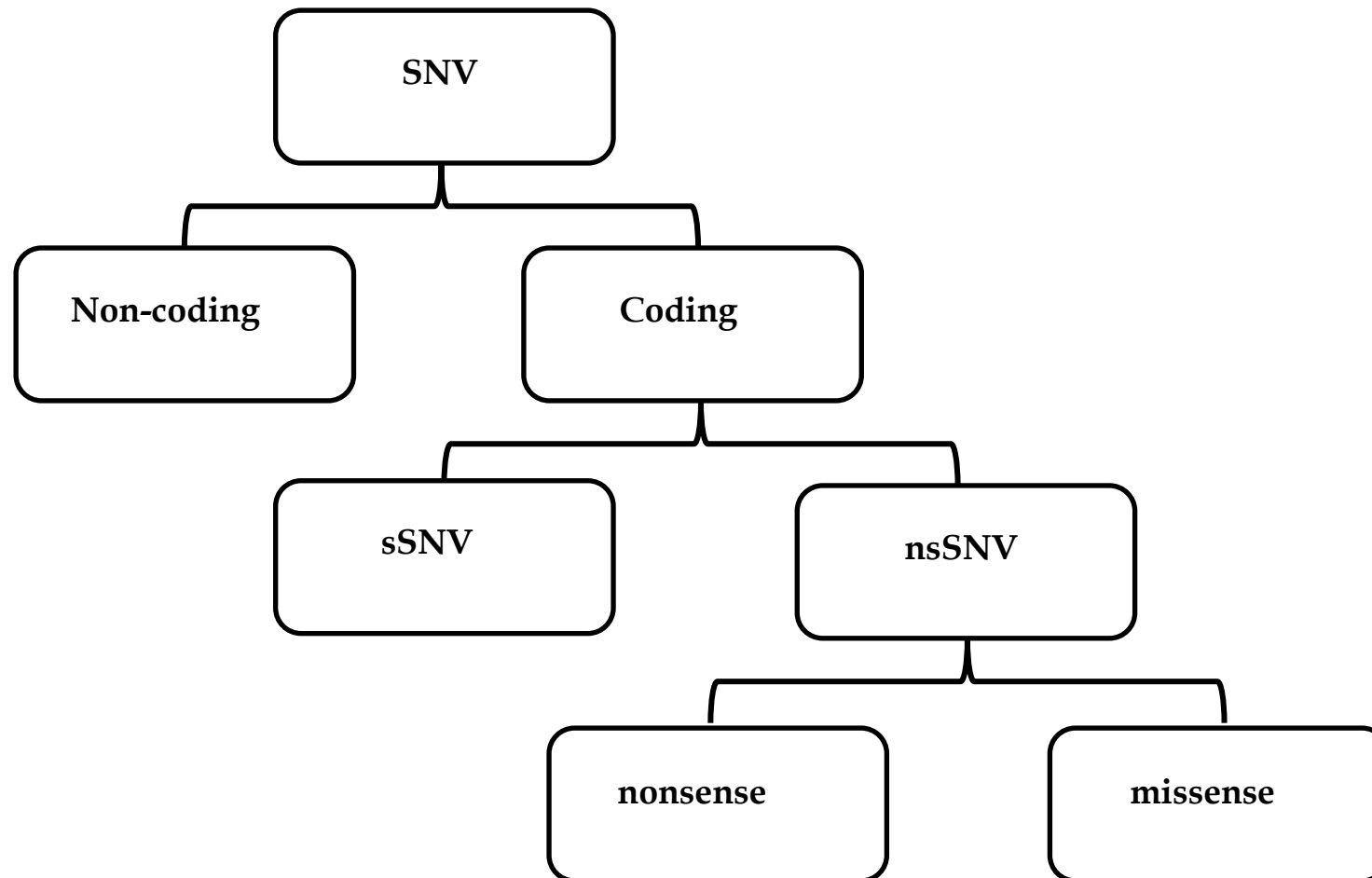


Figure 1.4 Types of single nucleotide variants. Where a single nucleotide variant (SNV) occurs will have an effect on the functional consequences of that variant. sSNV: Synonymous SNV, nsSNV: Nonsynonymous SNV.

Different from SNVs, deletions, insertions (including CNVs, cytogenetic indels, small indels), inversions and translocations can occur in different sizes within the genome, and again the effect will depend on where it occurs, whether it causes a frameshift in the triplet code or not and/or whether it affects the stability of the chromosome (e.g. the case of triplet nucleotide expansions and Fragile X syndrome [44]). For these DNA variations to be passed down the generations they have to occur in germ cells (i.e. sperm or an egg cell) as mutations which occur in somatic cells (e.g. liver, lung cells) are not passed onto the offspring. The latter type is not within the scope of this thesis thus will not be discussed further (for a comprehensive review on somatic mutations, see reference [45]).

Whole-exome sequencing (WES) will be used throughout this thesis. Therefore alongside non-exonic SNVs (including non-exonic SNPs), cytogenetic indels (i.e. large deletions, large insertions, duplications), translocations*, inversions†, large CNVs‡, whole chromosome losses/gains§ and variable tandem repeats** will also not be analysed and discussed [46]. WES can detect exonic SNVs (including exonic SNPs) and exonic indels of sizes ≤ 35 bases. This is due to the size of the reads (short length, ~50 bases) produced from the next-generation sequencing machines used in this thesis (detailed in section 1.6.2).

Similar to SNVs, small insertions and deletions (indels hereafter) can also have downstream effects on the translation of a protein – if in a coding region. This will depend on the size as well as the location of the indels (Figure 1.5). Definitions and examples are as follows:

* Rearrangement of parts between non-homologous chromosomes

† Chromosome rearrangement in which a segment of a chromosome is reversed end to end

‡ Copy number variations are alterations of long stretch of DNA sequence that results in the cell having a variation in the number of copies of that DNA sequence (or gene)

§ Most common extra autosomal chromosomes among live births are 21 (causes Down syndrome), 18 and 13

** Tandem repeats occur when a pattern of nucleotides is repeated and the repetitions are directly adjacent to each other

- 1- Frameshifting indels are indels with sizes which are not a multiple of three (e.g. 10, 20). These will cause a shift in the triplet code and therefore affect the remaining amino acid sequence starting from where the indel is located.

Example:

Wild type: GCT CTA TTA GGA GTT TGC TTA = ALLGVCL

2 base (TT) insertion: GCT CTA TTA **TTG** GAG TTT GCT TA- = ALL**LEFA**

- 2- Non-frameshifting indels are indels with sizes that are multiples of three (e.g. 6, 18). These types of indels will not cause a shift in the triplet code and therefore the remaining amino acid sequence will stay the same except for the indel itself.

Example:

Wild type: GCT CTA TTA GGA GTT TGC TTA = ALLGVCL

3-base (TTA) insertion: GCT CTA TTA **TTA** GGA GTT TGC TTA = ALL**L**GVCL

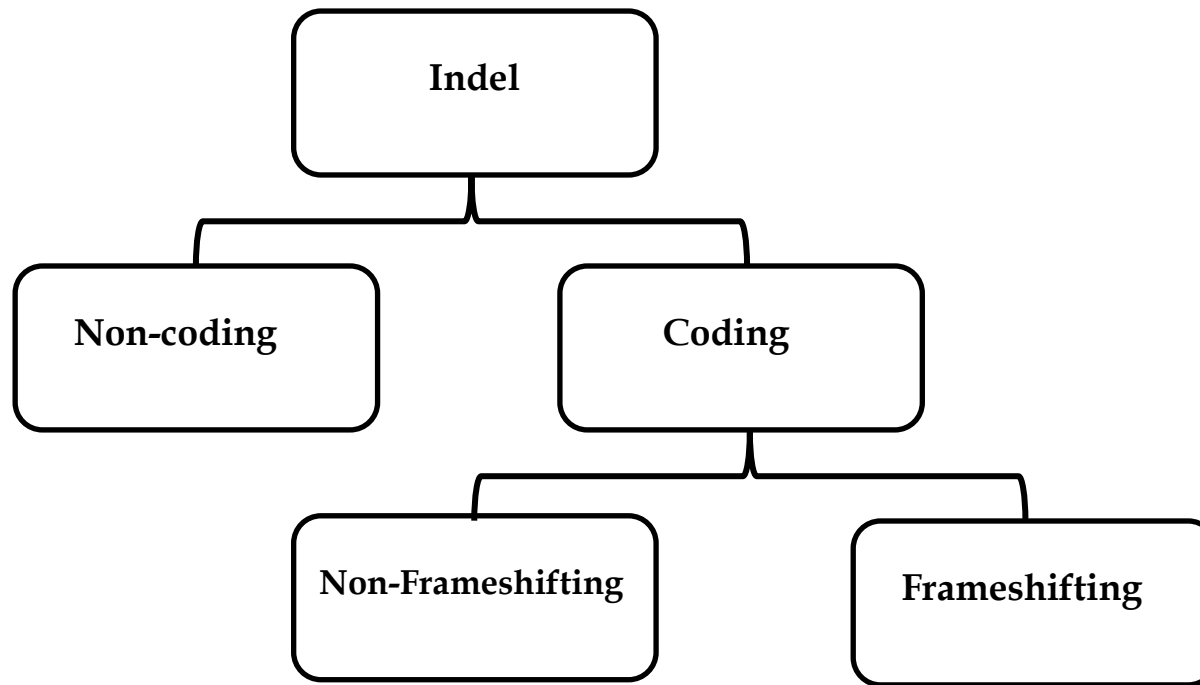


Figure 1.5 Types of indels. Where an indel occurs will have an impact on the functional consequences.

1.2. Epidemiology and Genetics

Epidemiology is the study of causes (determinants) and effects of clinical conditions in specified populations [47]. Epidemiological findings facilitate advances in evidence-based medicine and public health as well as informing policy decisions (e.g. study associating foetal alcohol exposure to lower IQ at age 8 [48]). Genetic research can contribute to the area and help epidemiologists by identifying, if applicable, genetic risk factors for any disease and/or health related trait (even behaviours such as alcohol and smoking addictions) and therefore providing targets for preventive and/or curative medicine.

Findings from genetic epidemiology studies have also been exploited by direct-to-consumer companies in order to make profit by capitalising on the average person's curiosity about their genomes. They genotype the DNA samples of customers to provide insights about their carrier status and disease risk for certain traits [49]. These reports can influence lifestyle choices and may incline an individual towards further preventive screening of potentially causal genes [50].

1.2.1. Genetic Epidemiology

Where applicable, genetic analyses of human disease can elucidate the mechanisms behind the observed symptoms; and provide targets for prevention and cure. A disease/trait is said to be monogenic when alterations in a single gene is sufficient for the phenotype to be expressed, given ordinary environmental (and genetic) background variation [51]. Due to negative selection against individuals with severe monogenic disorders, both for biological (e.g. reduced fertility and/or survival) and sociological reasons (e.g. stigma attached, not desirable to opposite sex), the frequency of monogenic disease causal variants are usually very low in the population (exceptions occur when balancing selection is observed, e.g. in the case of sickle cell anaemia and malaria resistance).

Monogenic disorders (also known as Mendelian disorders), because of their high penetrance*, show a distinct Mendelian pattern of inheritance such as autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive and Y-linked [51]; and knowing which inheritance pattern a disease follows is crucial in identifying any variant causal of a Mendelian disorder. This deterministic nature of the variants in monogenic disorders makes linkage/association analyses relatively easier to carry out; and this is reflected by the fact that the genetic cause of over 3000 Mendelian disorders has been determined [52]. However another 1500 (or more) Mendelian disorders are waiting for their causal gene(s) to be unearthed and also many of these disorders show complications such as allelic heterogeneity†, genetic heterogeneity‡, incomplete penetrance§ and pleiotropy** thus there is still a lot to be identified in this respect (see section 1.2.3 for more information on Mendelian disorders) [52].

Most of the common disorders such as cancers and diabetes (called complex disorders) however, do not show a follow a distinct pattern of inheritance due to their polygenic nature and tendency to be affected by environmental factors (e.g. diet, exercise, alcohol intake), thus dissecting causal variants of complex disorders is proving more difficult in relation to Mendelian disorders (see section 1.2.4).

With the ethical and health-related issues surrounding gene therapy techniques, it is currently not possible to cure a genetic condition through this route (barring a few exceptions, also except where interventions exist at the protein level); and therefore, as aforementioned, genetic epidemiology can provide a source of information for genetic counsellors to advise patients with the most up to date knowledge in genetics. If a clinically important genetic variant is found to be common in a population, then conventional epidemiologists can also play a part and influence

* The extent to which a particular gene or set of genes is expressed in the phenotypes of individuals carrying it, measured by the proportion of carriers showing the characteristic phenotype

† Phenomenon in which different mutations at the same locus causes a similar phenotype

‡ Phenomenon in which a single phenotype or genetic disorder may be caused by a multiple number of mutations

§ The probability of disease is less than 100% even when causal variant is present

** The phenomenon of one gene influencing multiple phenotypic traits

policy makers to provide funding into preventive interventions such as change of lifestyle for individuals who have a genetic predisposition to develop the common disorder, or directly into clinical trials for curative medicine (e.g. randomised controlled trials). Also making the public aware through strategies giving the message that ‘certain population groups and/or ethnicities may be more at risk for developing particular conditions than others’ may also encourage genetic screening amongst these ‘high risk’ populations. This is particularly important in populations such as Australia where 50% of the population are expected to be adversely affected by a genetic condition at some point during their lifetime and 28% of all infant deaths result from genetic factors [53]. Also the Ashkenazi Jewish community is well known to have elevated risks for a variety of disorders (e.g. Cystic Fibrosis, Tay-Sachs disease).

1.2.2. Terminology

Throughout this work, I will be using many terms to describe genetic variation and phenotypes. The term ‘variant’ will be used to encapsulate all types of genetic variation from a reference sequence whether it is a copy number variant (CNV), small insertion or a deletion event (indel), or a single nucleotide variation (SNV). The terms ‘single nucleotide polymorphism’ (SNP) and ‘common variant’ will be used interchangeably to describe SNVs with a minor allele frequency (MAF)* over 1% in a given population (and/or has global MAF >1%); and ‘rare variant’ will be used to describe SNVs with a MAF less than 1% (unless stated otherwise).

Individual variants will be described according to the Human Genome Variation Society (HGVS) mutation nomenclature (available at: <http://www.hgvs.org/mutnomen/>) and will be validated using the Mutalyzer online toolkit (available at: <https://www.mutalyzer.nl/>) [54]. The terms ‘disease’ and ‘disorder’ will be used interchangeably to describe severe deviations from the normal phenotypes (e.g.

* The frequency of the second most common allele at a loci where a SNP exists in a given population. The definition of ‘common’ varies amongst different projects and/or databases (e.g. >1% in 1000GP, >5% in some phases of the International HapMap project)

Cystic Fibrosis, Primary Ciliary Dyskinesia). The terms ‘mutation’ and ‘deleterious variant’ will be used to describe genetic variation which are known to be causal of a Mendelian disorder or have a large (highly-penetrant) effect on a common complex disorder (e.g. obesity, cancer). Although there are several definitions of a gene (see section 1.1.1), I will use the traditional definition of the term ‘gene’, which is used for regions of DNA which are known to code for a protein. The term ‘human genome’ will be used to describe all inherited DNA material found in the nucleus, thus will not include extranuclear DNA (i.e. the mitochondrial genome, mtDNA). Other genetic terms will be used in accordance with the Ensembl Variant Effect Predictor (VEP) sequence ontology (SO) terms – especially SO terms used within the ‘gene’ (orange sections, see Figure 1.6). Where necessary, rare stop gains/losses, start losses, splice-site acceptor/donor variants, missense mutations and exonic indels will be pooled under the name ‘ Φ ’ (capital phi, acronym for ‘predicted high impact’) mutations (see section 9.8 for details).

The term ‘incest’ will be used for first degree unions (e.g. Father-daughter union, sibling unions), whereas ‘consanguineous’ unions will refer to third degree unions (e.g. first cousin unions, double first cousin unions) and unions between more distantly related individuals (e.g. second cousin unions) which would still be classified ‘consanguineous’. Additionally, all types of consanguineous unions are depicted in section 9.7 (in the appendices chapter).

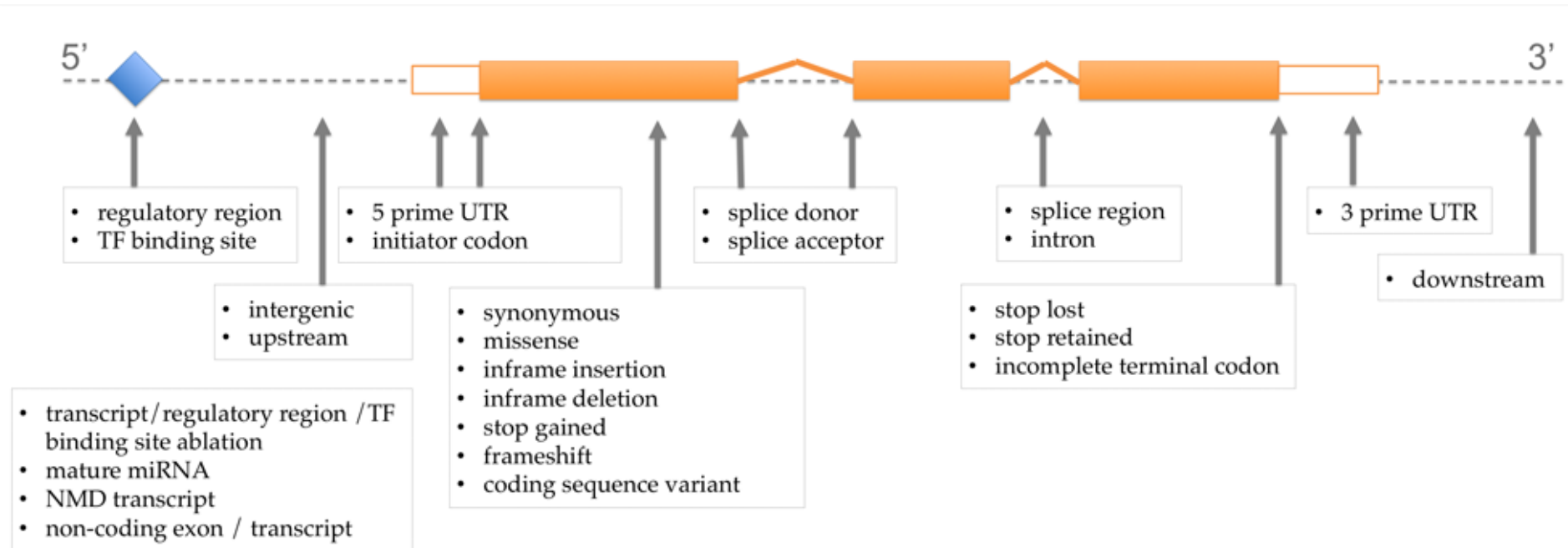


Figure 1.6 Sequence ontology terms from the Ensembl Variant Effect Predictor. These terms are used to describe where a variant has occurred and its consequences to the transcript. More information can be found on the Ensembl Variant Effect Predictor website (see http://www.ensembl.org/info/genome/variation/predicted_data.html). Figure reproduced with permission from Ensembl (open source, version 73) [55].

1.2.3. Mendelian disorders

As briefly described above (in section 1.2.1), Mendelian disorders are monogenic thus the inactivation of a single gene or the malfunction of the gene product (i.e. protein) is sufficient for the disease phenotypes to be expressed. The term ‘Mendelian’ comes from an Austrian (and Augustinian) monk and later-turned-botanist Gregor Johann Mendel (born 1822) who was the first to empirically show that the effects of inheriting an allele on a single chromosome (i.e. heterozygous) can be different from when the same allele is present on both chromosomes (i.e. homozygous). Many proteins, especially enzymes are produced in adequate quantities even if one copy of the gene is inactivated by a loss of function* mutation which can either halt protein production completely (e.g. through nonsense mediated decay, NMD hereafter [56]) or by the production of a malfunctioning protein. Complete absence of a correctly functioning gene product due to loss of function mutations in a homozygous state will cause a deficiency of the protein, and this will usually cause disease due to downstream effects†.

Clinical diseases caused due to complete lack of gene product are usually very rare (or absent) in an outbreeding population as the disease causal alleles are almost always in a heterozygous form – thus many of them can stay in the population for hundreds of years silently (i.e. without causing disease) [57, 58]. The total number of (characterised and suspected) monogenic diseases is thought to exceed five thousand [52, 59]. Adding to this colossal figure, an increasing proportion of common diseases previously thought to follow a complex multifactorial inheritance pattern (e.g. schizophrenia, autism, intellectual disability; see Chapter 4 for details on the latter disorder), are now believed to represent heterogeneous collections of monogenic disorders [59, 60]. Although everyone is a carrier of a combination of several loss of function and/or autosomal recessive mutations in certain genes,

* All types of variation (e.g. SNV, deletion) which cause the gene product to be dysfunctional or be completely removed (e.g. due complete deletion of gene, nonsense mediated decay)

† Although see ‘analbuminaemia’ in the literature for an interesting/unusual example

usually two unrelated individuals (who will go on to have offspring) will not be carriers for the same mutation. Therefore homozygous forms of rare recessive disorders are almost always found in the offspring of consanguineous unions (details in section 1.3), in endogamous populations and/or in isolated populations with small population sizes.

Although these types of mutations are broadly called 'recessive' or loss of function, they can follow different patterns of inheritance depending on which chromosomes they reside on. As there are two copies of each autosome (2 x chromosome 1 to 22, under normal conditions), two copies of a recessive mutation are also needed to express the disorder (Figure 1.7). These types of disorders are called 'autosomal recessive' disorders (e.g. Cystic fibrosis, thalassaemia, Tay-Sachs disease, Primary Ciliary Dyskinesia – see Chapter 3 for details on the latter). If both the father and the mother are carriers of an autosomal recessive mutation, then there is a 25% chance that the offspring will be affected. Some of these disorders are more common in individuals of certain ethnic backgrounds due to endogamy and/or consanguinity (details in section 1.3). Although most affected individuals will be homozygotes for the causal variant, sometimes compound heterozygotes* are also observed amongst individuals with autosomal recessive disorders.

If recessive mutations reside on sex chromosomes, the prevalence of their corresponding disorder differs according to sex. Since virtually all real life examples of sex linked recessive mutations are located on the X chromosome, the 'theoretical' Y chromosome linked mode of inheritance will not be discussed. As there are two copies of the X chromosome in females, it is very rare that X-linked recessive disorders are expressed in females as it requires both the father (who would be affected) and the mother to be carriers. Therefore, although recessive, since the males only have one copy of the X chromosome, even a single copy of the mutation is

* The condition of having two heterogeneous recessive alleles at a particular locus that can cause genetic disease – ultimately causing disease

enough for the disorder to be expressed – since they will be hemizygotes*. If the mother is a carrier of an X-linked recessive mutation, then there is a 50% chance of her son being affected, whereas the females are only at risk if both parents possess the mutation. Examples of X-linked recessive disorders include red-green colour blindness† (see reference [61]), Duchenne muscular dystrophy and Haemophilia A.

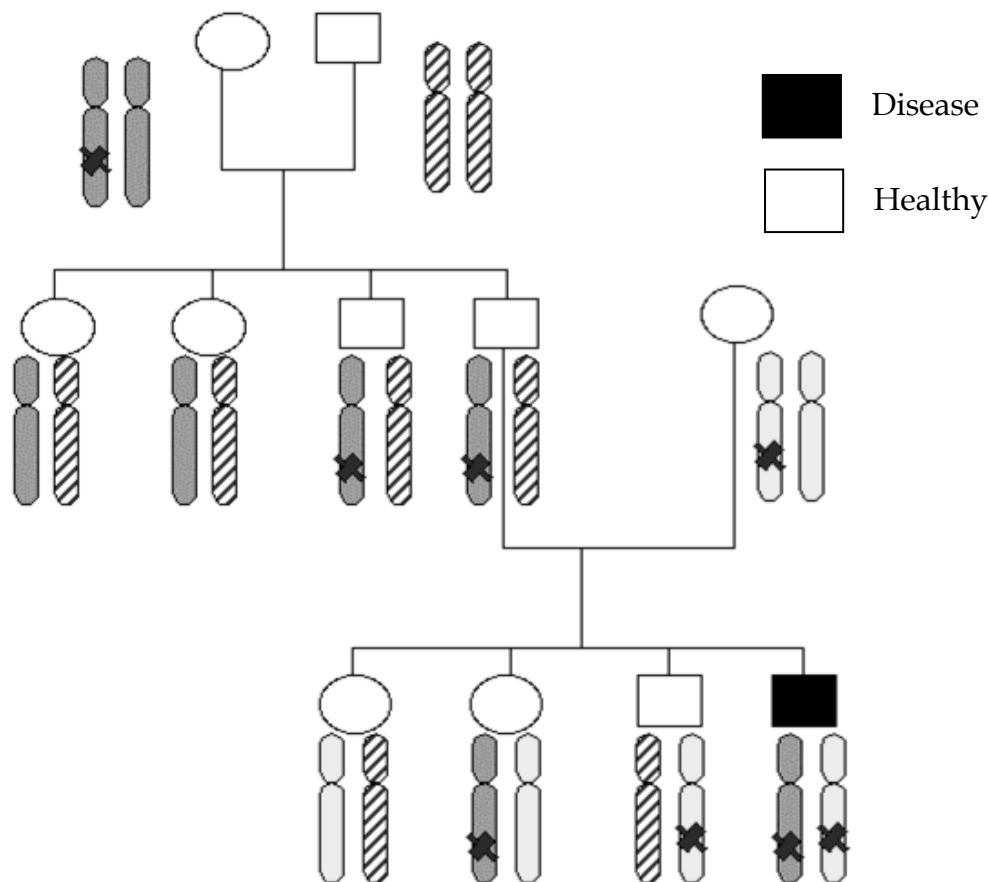


Figure 1.7 Autosomal recessive pattern of inheritance: Both of the affected male's parents are carriers of the recessive mutation (X represents recessive mutation). For autosomal recessive mutations with a low frequency, it would be extremely unlikely to see affected individuals, unless the parents are consanguineous (discussed in section 1.3). The figure is a simplified depiction of what occurs in reality as meiotic recombination events have been ignored – see Figure 1.11 for a more realistic depiction.

* The state of having one or more genes (as in a genetic deficiency or in an X chromosome paired with a Y chromosome) that have no allelic counterparts

† A fitting example showing gender bias - as up to 8% of males are affected, whereas the prevalence is below 1% in females

Other Mendelian modes of inheritance include 'autosomal dominant' and 'X-linked dominant'. Early-onset (i.e. severe before adulthood) forms of these types of disorders are usually caused by *de novo* mutations which occur in the germ cell of one of the parents (i.e. sporadic, thus the family will usually have no history of the disorder). If a disease is autosomal dominant, it means that a single mutation is sufficient for the expression of the disease phenotypes (Figure 1.8). Detrimental early-onset forms of these disorders cannot be passed down the generations thus relatively common autosomal dominant disorders are mostly in the form of late-onset disorders (e.g. Huntington's disease [62]), with no effect on fertility during early adulthood (e.g. Neurofibromatosis-1, Achondroplasia [63]).

Similar to the autosomal version, only a single mutation is required for X-linked dominant disorders. In this case, in contrast to X linked recessive, there is no gender bias as both sexes have a 50% of inheriting the mutation from their affected mothers. If the father is the only affected parent, then there is no chance of the son being affected (barring an unlikely event of a *de novo* mutation in the germ cells of the mother). Famous examples of X-linked dominant disorders are Charcot-Marie-Tooth disease, Incontinentia pigmenti and X-linked hypophosphataemia [64-66].

Mitochondrial (genome) diseases also follow a distinct pattern of inheritance* but are not counted as a Mendelian disorder – due to the mitochondrial genome being located outside of the nucleus. They are a group of disorders affecting about 1 in 8000 in the population caused by dysfunctional mitochondria due to mutations in the mitochondrial genome (usually abbreviated mtDNA) [5].

Further to all mentioned before, many Mendelian disorders (especially rare ones) have no therapy let alone a cure, thus prevention through genetic counselling is crucial. Economics and profit plays a huge part in the development of treatments/cures, and rare diseases are not attractive to pharmaceutical companies in this respect. The recently developed CRISPR/Cas9 technique is showing some

* Mother passes mitochondrial DNA to all children, but then only female children passes onto their offspring

promise, however it is still early days to discuss the efficacy of any preventive programme [67].

As autosomal recessive mutations account for the main differences in relation to genetic disorders between consanguineous and outbred populations (coined the term inbreeding depression, details in section 1.3), they will be the main focus of this thesis; and other forms of Mendelian diseases will be discussed no more.

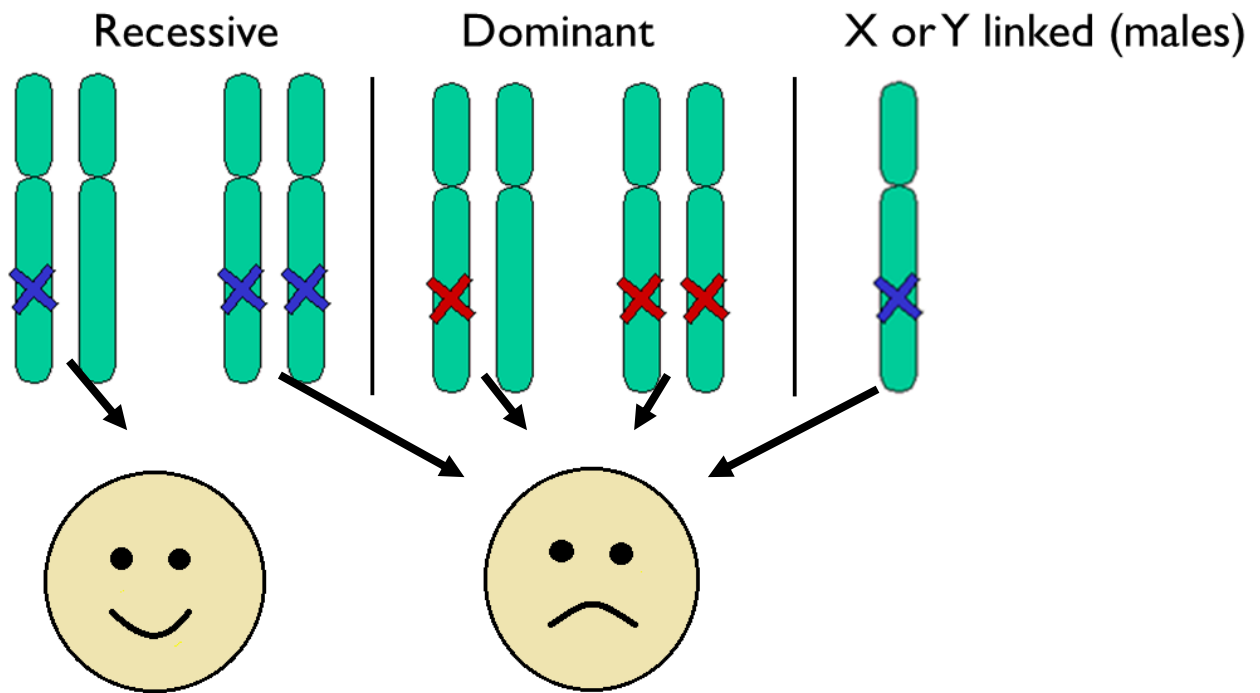


Figure 1.8 Contrast between recessive and dominant mutations. Which chromosome a mutation occurs is also important (i.e. autosome or sex chromosome). Blue X's depict recessive mutations and Red X's depict dominant mutations. Smiling face indicates no disease; and sad face indicates disease phenotypes.

1.2.4. Past and present hypotheses on Mendelian disorders

Within-family linkage studies have led to the identification of many highly-penetrant disease causal loci and served human disease genetics greatly. As the

name suggests, these studies capitalised on the phenomenon of linkage* and made use of genetic markers such as short tandem repeats (STR), SNPs and restriction fragment length polymorphisms (RFLP) scattered across the genome. The aim would then be to identify a marker which co-segregates with the disease within the family; and this marker would then give valuable information about which gene(s) may be causal†. Logarithm of the odds (LOD) scores‡ were also being used within the field by statistical geneticists to help map disease loci where inadequate number of affected offspring were present within families in order to combine marker information from multiple families (or large pedigrees). Lander and Botstein introduced a powerful method called ‘homozygosity mapping’ to identify disease causal regions in the offspring of consanguineous families [68]. Their strategy made use of the fact that in consanguineous families where a recessive disease is present, a region of many centimorgans§ (cM) spanning the disease locus is almost always ‘homozygous by descent’ in the affected offspring (termed ‘autozygous’ today). Although they required a complete RFLP linkage map in their 1987 paper, the underlying idea of using ‘identical by descent’ regions to map disease loci is still being used today.

These studies helped identify causal regions/genes, however they lacked the resolution needed to identify the causal variant itself. Advancements in sequencing (e.g. Sanger sequencing, NGS) would be required to obtain the sequence of the identified region(s) and lead to the identification of disease causal mutations. Also these traditional methods struggle with genetically heterogeneous diseases as they do not have the power to detect rare disease causal variants (i.e. to reach a LOD score of 3). Therefore identifying the causal variant even for common complex

* Two genetic loci are ‘in linkage’ if the alleles at these loci co-segregate more often than that would be expected by chance (i.e. >50%). This is due to the two loci being very close to each other on the same chromosome and therefore there is a smaller probability of a recombination occurring between them during meiosis

† And also which individuals may be carriers and/or at risk of expressing disease

‡ To the base 10, thus a LOD score of three translated to a probability of 1/1000

§ One centimorgan is defined as the distance between loci/markers for which the expected (average) number of recombination events in a single generation is 1/100

diseases such as schizophrenia, intellectual disability and autism would be an unfeasible task (let alone rare and complex disease).

Today, traditional techniques have been improved with recent technologies such as dense SNP chip arrays and NGS (details in section 1.6), and the availability of large mutation databases. WES has also changed the landscape of Mendelian genetics as researchers are now able to identify highly-penetrant disease causal mutations with relatively small number of affected individuals, sometimes even a single proband may be sufficient (details on WES in section 1.5).

1.2.5. Complex disorders

As briefly mentioned before (in section 1.2.1), complex disorders are multifactorial thus are more enigmatic in relation to Mendelian disorders. Common complex disorders such as cancer, obesity and type-2 diabetes affect millions of individuals worldwide; in fact there are 2.9 million people affected by the latter disorder in the UK alone (and around a million are thought to be undiagnosed) [69]. As they are examples of common disorders requiring constant therapy and/or care, they account for a large proportion of the state health services budget. For example obesity related health problems cost more than £5 billion to the National Health Service (NHS); and likewise it is estimated that around 10% of the NHS' budget is being spent on diabetes [70, 71]. Furthermore, coronary heart disease, another common complex disorder, is the leading cause of mortality in the industrialised world [72].

Complex disorders' genetic makeup is not fully understood but it is thought that they can be influenced by a combination of rare and common variants with ranging (between high and low) effect sizes (penetrance). Environmental factors can also play a significant role and the proportion of variance explained by the environment varies widely depending on the disorder. Thus individuals with a 'risky' genetic background (i.e. predisposition) for the disease may turn out to be perfectly healthy

with an appropriate lifestyle*, whereas individuals without too much genetic risk can contract the disease (due to extreme exposure to environmental factors/determinants such as smoking, stress and/or alcohol); and this is why complex disorders do not follow a distinct Mendelian pattern and thus are much harder to dissect.

Complex traits/disorders puzzled many late 19th and early 20th century scientists. Francis Galton published his 'blending' characteristics (what is now called quantitative traits) theory in 1885 and then a refined version in 1897 [73]. Ronald Fisher then published a landmark paper in 1918 which showed that common complex disorders and quantitative traits (e.g. height) could also be explained in the context of Mendel's laws as they could be caused by variation at many different loci [74-76]. However, although Mendelian and Galtonian theory can correctly explain many (Mendelian and complex) human diseases and traits, they do not describe and/or fit all real-life observations [72]. Therefore alleles with varying penetrance, interaction between genetic loci (i.e. epistasis) and the effect the environment can have on complex traits (i.e. epigenetic factors) were added to more refined complex trait genetics models in subsequent years. These new models are now being tested through functional and molecular studies, genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS). Figure 1.9 represents the relationship between penetrance of disease risk alleles and their frequency, and how different combinations of these two factors affect their discoverability in genetic association studies.

GWASs have identified many associations between common variants and complex traits [77, 78]. However, in contrast to analysing a Mendelian disorder (where a small number of participants, sometimes even a single individual's WES data can be sufficient [79, 80]), carrying out a well-powered GWAS requires thousands of participants' (cases and controls) phenotype and genotype data, which can cost a fortune; and to make matters worse, in contrast to previous expectations (which

* And luck; or as some like to put it favourable stochastic events

were raised by the proponents of the common disease common variant (CDCV) hypothesis – see section 1.2.6), they have failed to explain most of the heritability of these common complex traits (coined the name ‘missing heritability problem’ by Manolio *et al* [77]) – thus they have limited value for disease prediction with odds ratios rarely above 1.3 for a given SNP reported in the literature [40, 81, 82]. The onus is now on rare variants to explain these gaps in heritability of many of the common complex disorders. Some have even suggested that many of the GWAS signals could be ‘synthetic associations’ caused due to ‘tagging’ of one or more rare variants by common variants [83].

The debates on what the ‘true genetic mechanism’ behind common and complex disorders is centred around three models at present [84]. These models are the (i) rare allele model, where rare alleles with large effects constitute the genetic basis of common disorders (ii) infinitesimal model, where thousands of alleles with small effects are responsible for the disease phenotypes; and (iii) broad sense heritability model, where gene-gene (GxG) interactions and gene-environment (GxE) interactions contribute to disease in a non-additive way – besides the effects of rare and common alleles [84]. In cases where the outcome is binary and not continuous (e.g. disorder or not, discontinuous multifactorial traits), identifying the right genetic model behind each disorder is crucial, making disease risk prediction a bit more of a challenging process as the multitude of factors may or may not exceed the ‘liability threshold’ as proposed by Mossey [72, 85]. The evidence behind each model is briefly discussed in the next section (for a comprehensive review on arguments for and against the role of both common and rare variants on common disorders, see reference [84]).

At current, the effect of consanguinity *per se* on common complex disorders is very poorly understood [86], but it is bound to become clearer as a consensus on the right models is reached (see section 1.3 for a literature review on consanguinity’s effect *per se*).

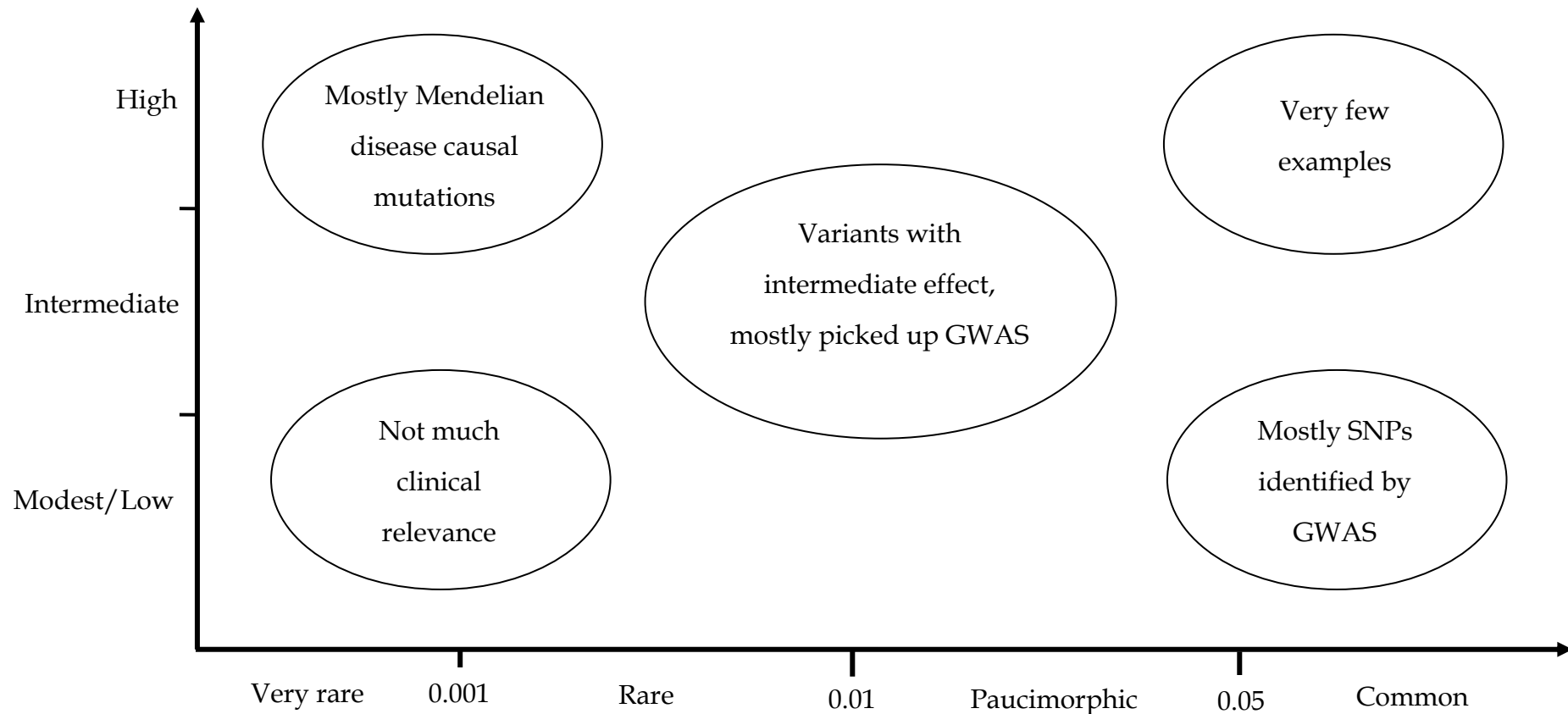


Figure 1.9 Relationship between allele frequencies, penetrance and discovery: Highly penetrant mutations are expected to be low under normal circumstances, unless balancing selection acts upon the mutation (e.g. the case of sickle cell anaemia and malaria resistance). Rare variants with modest/low effects are very hard to identify genetically and they do not have much, if any, clinical relevance at a public level. There are very few examples of common alleles with high penetrance on a multifactorial disorder (e.g. $\epsilon 4$ allele at the *APOE* locus which is associated with Alzheimer's disease). The main focus in complex trait genetic studies at present is to identify as many variants with intermediate or high penetrance as possible however rare they may be – which will fill in some of the gaps in heritability. “Pauci” means ‘few’ and is used in this context to describe frequencies between ‘common’ and ‘rare’ [87, 88]. Image adapted from reference [89].

1.2.6. Past and present hypotheses on Complex disorders

Once the inheritance pattern of a Mendelian disorder is figured out, the next task is to identify rare variants which fit the model; and with the availability of sufficient and reliable sequencing data this job becomes relatively straightforward. However, understanding the genetic makeup of common complex disorders is a long winded task even with all the technological advances. A large proportion of the heritability of these disorders, still remain unexplained.

In 1996, the CDCV theory was published, proposing that the combination of effects of common genetic variants with intermediate effects was to blame for common complex disorders [90, 91]. The prime example used by the proponents of the CDCV hypothesis was the $\epsilon 4$ allele* in the *APOE* gene which dramatically increases the risk of Alzheimer's disease (and heart diseases) and had a population frequency ranging from 5.2% to 40.7% depending on the population [92, 93]. Many highly powered GWASs were carried out consequently, hoping to identify common variants which would help predict disease risk for each individual. However GWASs proved to be the 'end' of the CDCV hypothesis as results showed that, completely contrary to initial thoughts, the heritability accounted by common variants remains modest (less than 5% in most cases) even when the effect of all common variants were summed [84]. Building on this setback for the CDCV model, proponents of the rare allele model began to argue whether common variants have any effect at all, as the associations reported could be 'synthetic' due to tagging of rare alleles (with large effect on the disorder) in LD with the 'tag' SNP which represents the region in the commercial SNP arrays [83].

For this reason alternative models were proposed, one of which is the common disease rare variant (CDRV) hypothesis [94, 95]. It is derived from the assumption

* Homozygous T>C change at **both** rs429358 and rs7412 loci

that common variants must have arisen many generations ago and because they do not reduce the 'fitness' (e.g. survival, fertility) of an individual significantly, if at all, they have become common as they have not been selected against by natural selection [94]. Mutations causing Mendelian disorders which affect fitness detrimentally will always remain rare in the population exactly for this reason (i.e. due to their high penetrance) however multiple rare variants (which arose recently) with intermediate effects (i.e. with higher influence on complex disorder than common variants and lower penetrance than Mendelian disorder causal mutations, see Figure 1.9), may have remained which is thought to explain most of the missing heritability that is observed when only the effect of common variants are taken into account. Balancing selection can also cause deleterious variants to remain at the population level due to heterozygote advantage over homozygotes*. One classical example is the advantage of being a privileged heterozygote† for malaria resistance. This is a well-known and understood example whereby heterozygotes at the sickle cell anaemia causal site of the haemoglobin beta gene (*HBB*) have increased resistance to malaria in comparison to homozygotes [96].

Many rare and highly penetrant variants (e.g. loss of function variants) exist in the human genome, however since they do not have any detrimental effects to the fitness (i.e. health and fertility-wise) of an individual until later in life (if at all), they are not removed from the population [97]. This is because most occur in non-essential genes such as olfactory, taste and hearing-related receptors [97], nevertheless these variants still being relatively rare ($\leq 5\%$) indicates that natural selection does select against them although not as harshly as it does against Mendelian disorder causal mutations. The frequency of alleles which are rare in a large population, can rapidly increase in (or be removed from) a population due to random genetic drift, bottlenecks (e.g. due to catastrophes) and/or founder effects – even if it is a non-consanguineous population (see section 1.3 for effect of consanguinity and endogamy on a population). Illuminating the role rare variants

* Coined the term 'overdominance'

† A term I came up with to describe individuals who are heterozygotes for a balancing mutation

has on complex disorders requires a better understanding of the mechanisms involved as well as more reliable sequencing techniques and larger sample sizes.

A thorough analysis of *de novo* mutations should be next in order, which remained an unexplored area until very recent [98]. *De novo* mutations are demonstrated by Roach *et al* (and later replicated by Conrad and colleagues) to occur at a rate between 1.1×10^{-8} to 3×10^{-8} per base per generation [29, 30]; and this figure can also be used as an estimate of the probability of a deleterious mutation occurring in any individual*. As aforementioned in section 1.2.4, the biggest challenge when analysing complex disorders is reaching a ‘good enough’ sample size (which may require delicate power calculations, see reference [99] for details); and with rare variants, this task becomes even more challenging statistically, analytically and cost-wise. It is realised that traditional techniques will not work efficiently (due to low statistical power of the methods) and new tests and methods are being developed such as burden tests [100, 101] which involve collapsing of rare variants which occur in the same gene or regions of interest. There are certain assumptions that each burden test relies upon, and which may be false; but these will not be discussed here as these issues are outside the scope of this work.

However other ideas and hypotheses have also been shared and proposed to explain the missing heritability problem that is observed with many complex disorders – although this notion is being challenged by some studies suggesting that the heritability isn’t missing, it is just spread across a very large number of variants of tiny effect [77, 102]. Epistasis (gene-gene interactions)[†], gene-environment interactions and epigenetic[‡] factors have been proposed as the missing links (as part of the broad sense heritability model [84]), whereas others have claimed that the heritability estimates are inaccurate and the genetic component of many complex

* Although some regions in the genome are more mutable than others

[†] The interaction of genes (that are not alleles), in particular the suppression of the effect of one such gene by another

[‡] Refers to functional modifications to the genome that do not involve a change in the DNA sequence. Examples are DNA methylation, acetylation and histone modifications, which all serve to regulate gene expression

disorders has been overestimated. It should be noted that epistatic interactions between the mtDNA and nuclear loci is also far from being completely understood thus should also be paid more attention [103]. It may be that all of the above-proposed models are correct in a sense that they could be contributing to different complex disorders in some way or another*. Since our understanding of the interactome (i.e. interactions between genes and their products) is far from complete, results from genetic association studies should not be taken as ‘truth’ before confirmation through functional and/or molecular studies. The wider use of WES and WGS in the near future is also expected to contribute greatly to our understanding.

For these hypotheses to be analysed in the most appropriate way, larger epidemiological studies have to be initiated with components from all parts of the ‘omics’ spectrum such as epigenomics, transcriptomics and metabolomics, in addition to the traditional genetics/genomics approaches. As this thesis mainly focuses on consanguinity, discussions on the broader issues of Mendelian and common complex disorders will be discussed no further. Rather specifically, the effect of consanguinity on both of these types of disorders will be presented and discussed in the next section (section 1.3).

1.3. Consanguinity and Genetic research

Inbreeding, using all sorts of model organisms from fruit flies (*Drosophila melanogaster*) to baker’s yeast (*Saccharomyces cerevisiae*), Arabidopsis (*Arabidopsis thaliana*, a small flowering plant) to the house mouse (*Mus musculus*), has served the area of genetics greatly as ‘a whole’ alongside facilitating the understanding of our own genome. The homozygous (or nullizygous†) effect of a mutation/allele (e.g. gene knockout, gene knock-in) is studied by mating two very closely related parents

* Gibson has listed twenty arguments for and against both rare and common allele models (see reference 84)

† The state of carrying two loss-of-function (or ‘null’) alleles i.e. both copies of gene are dysfunctional

(usually sibling pairs) and the resulting phenotype is assessed in the offspring. This then gives great insight into the functions of the gene as well as the pathways its product (i.e. protein) is involved in. However due to ethical and legal reasons, these studies cannot be carried out with humans; and this factor, but rightly so, leads to progress in human genetics being much slower in relation to progress made in model organisms. This is where studying consanguineous populations can speed things up as consanguineous unions, albeit still nowhere near, are the closest humans get to the level of inbreeding which occurs in the abovementioned genetic studies of model organisms*. Consanguinity can be described as “kinship characterised by the sharing of common ancestors” [104], thus the term does not include affinal (i.e. related by marriage, e.g. cousins in law, step cousins) or fictive (i.e. sociologically kin-like relationships such as relatives by adoption, wives of a pair of brothers and godfathers) kin. Homosexual unions are obviously not included amongst the terms ‘consanguineous’ unions as there can be no biological relationship between the parents and their children (at least with today’s technology), assuming there is (at least) one.

In clinical/medical genetics, any union between individuals who are related as second cousins or closer is considered consanguineous† (details in section 1.3.4). From a statistical genetics perspective, although consanguinity does not alter the frequency of alleles directly, it affects genotype frequencies as the violation of HWE assumption of random outbreeding means that, for a given allele frequency, the genotypic frequencies will differ from those expected. Therefore consanguinity can affect the prevalence of genetic disorders. Consanguineous relatedness is defined according to the likelihood of sharing similar genetic makeup with respect to common ancestors. Just as parents and children, a pair of brothers (or sisters) shares approximately half‡ of their chromosomal constitution. These relationships are termed ‘consanguineous of the first degree’. Using the same principle, an aunt or

* Which sometimes reaches figures close to a 100%

† Somewhat arbitrarily

‡ The term ‘approximately half’ is used because of the chance factor during recombination in meiosis – process which produces the haploid sperm and eggs

uncle shares about quarter of the genetic makeup with their niece/nephew (as do half-siblings) thus are called second degree relatives; and following this logic, first cousins are consanguineous kin of the third degree – the most common of type consanguineous unions [105, 106].

A loss of function (e.g. autosomal recessive disorder causal) allele may be rare in a population, but once the allele has been inherited by the parent (or occurred *de novo*), it has the same probability of being passed onto the child as any other common* and/or neutral allele. The chance of the offspring being homozygous for that allele would still be extremely low as the chances for the other parent to be a carrier of the same allele are very low. However this probability of ‘both parents being carriers’ is at its greatest in the case of consanguineous unions as they share a relatively close common ancestor who is more likely to pass on his/her mutations to the parents. Genetic effects of consanguinity is most appreciable in rare autosomal recessive disorders [107, 108]; rarer the prevalence of a disorder, higher the proportion of offspring with consanguineous parents in relation to total of affected individuals. This generalisation can be applied to all autosomal recessive disorders, whether they are truly monogenic (e.g. Cystic Fibrosis) or genetically heterogeneous (e.g. Primary ciliary dyskinesia, see Chapter 3 for details).

The subsequent sections will introduce and briefly discuss the effects of consanguinity on human populations, the studies carried out on the effect of consanguinity *per se* on different types of traits and disorders, and statistics on consanguinity in different world populations and their implications. A historical perspective on consanguinity and incest will also be included to provide a different viewpoint to today’s negative outlook towards these phenomena in the Western world.

* According to the infinite-site model, even homozygosity for common alleles may be due to distant relatedness due to the low probability of the same mutation occurring independently in different individuals

1.3.1. Consanguineous societies and genetic disease

Intra-community marriage (i.e. endogamy) rates are high in regions where consanguineous marriages are favoured maybe due to influence of one's culture on adjacent communities (coined the term 'transculturation' by Fernando Ortiz [109]). Evidence for this comes from non-Muslim populations living in the Arab world such as the Lebanese [110], Palestinian [111] and Jordanian [112] Christian populations, thus consanguinity in these parts of the world is not only confined to Arab Muslim communities* [113]. Endogamy can occur at different levels such as within clan/tribe which is common in Arab societies (e.g. hamula or kabeela [113]), within the same caste which is common in India; and/or within biraderi (i.e. patrilineage) which is common in Pakistan [86]. Due to lack of gene flow from other communities, unequal distribution (and therefore clustering) of founder mutations† can occur as a consequence; and the effects of these founder mutations and/or genetic drift‡ can come into the fore which can cause adjacent villages and even sub-communities living within the same town to exhibit very different (genetic) disease profiles. Under these circumstances, a disease causal mutation can rapidly increase in frequency. In fact this is shown to be exactly the case in a variety of tribes and villages where there is clear separation of diseases [86, 114, 115]. Allelic heterogeneity in individuals with very rare autosomal recessive disorders (e.g. Alström syndrome) has also been observed in a number of highly consanguineous populations such as the Arab and Druze communities [116, 117]. This phenomenon can be ascribed to random chance (of mutational events) or selective heterozygote advantage [105, 116, 117], with the latter hypothesis requiring further research. Real life examples of clustering of certain disease can be found in some Arab societies (coined the name 'Arab Genetic diseases' by [118]) such as Bardet-Biedl syndrome (an example of a ciliopathy as discussed in section 4.1), autosomal recessive severe

* It must be mentioned however that them being Muslim *per se* has got no direct influence on choosing to engage in a consanguineous union – which will be discussed later

† They are disease causal mutations which were present in the genome of one or more individuals who are the founders of a distinct population (usually with small starting sizes and without much admixture in future generations, resulting in endogamy/consanguinity)

‡ It is the cause of drastic change in the frequency of an allele in a population due to random sampling

childhood muscular dystrophy and osteopetrosis [118]. Diagnosis of a (very) rare recessive disorder caused by the same mutation in several (seemingly) unrelated families is indicative of an old founder mutation (and the age of the allele will depend on how far related these families are), whereas presence of these types of disorders in several members of the same family is indicative of a recent founder mutation (e.g. tribal/clan founders). Recessive alleles can also be introduced (i.e. exported) into a gene pool by admixture* through unions from another community, which can then increase in frequency through endogamy and/or consanguinity [119].

1.3.2. Inbreeding depression in humans?

The phenomenon of inbreeding depression† has a long history of extensive studies in animals and more recently in humans with many underlying mechanisms suggested‡ (e.g. increased homozygosity at loci with overdominance i.e. being a heterozygote is advantageous over homozygotes, expression of recessive deleterious alleles, inbreeding-environment interactions) [120], with latest theories involving epigenetic mechanisms (e.g. DNA methylation, histone modifications, RNA interference) [121]. Continually, increased performance of the first filial generation (F₁) has been presented in plant studies, where hybrids often have higher quality than their inbred parents (i.e. heterosis) [120, 122]. Although the exact nature of the relationship between human inbreeding and fitness is still not completely understood [123], analysis of the genetic load§ in consanguineous populations have shown inbreeding's effect to be moderate in humans [105]. Mouse models have yielded interesting but contradicting results, where healthier offspring compared to the F₁ generation were produced through inbreeding of knock-out (for cytochrome

* It occurs when individuals from two or more previously separated populations begin interbreeding; and results in the introduction of new genetic lineages into a population

† Deleterious effects (e.g. reduced survival and fertility compared to offspring of unrelated individuals) expressed in the offspring due to unions amongst closely related individuals

‡ See reference ¹²⁰ for a comprehensive review on the genetic basis of inbreeding depression.

§ Reduction in fitness due to deleterious alleles maintained within a population

P450 genes in the *Cyp1*, *Cyp2*, *Cyp3* and *Cyp4* gene families) mouse lines – which has introduced a new term and paradigm into the field: ‘inbreeding de-repression’ [124].

Human population based studies analysing the effect of consanguinity *per se* on mortality and childhood defects reported an excess of 20% to 35% in the offspring of incestuous unions (of the first degree) compared to statistics from populations where consanguinity levels are low (e.g. in Western countries) [104]. This increase in risk is relative to the worldwide figures of 3% to 5% of all live newborns which have a clinically significant birth defect [113]. The risk for the offspring of double first cousins is estimated to be three times higher than the population background risk [125]; and this effect is significantly lowered as the degree of the consanguineous union decreases, with the increase in observed (infant and/or pre-reproductive) mortality and additional birth defects being between 3% and 5% (with 3.5% reported in the largest study comprising of over 2 million individuals from 69 populations [86], $r^2 = 0.70$, $P\text{-value} < 10^{-5}$), and 0.7% to 7.5% percent (both figures vary between populations [126]) respectively for offspring of first cousin unions [86, 108, 113, 127-129]. These conflicting results reflect the lack of standardisation in methods and failure to control for the varying socio-economic and environmental factors in each study [86, 105]. These values may also be confounded by other non-genetic influences such as the systematic differences in definition of a ‘birth defect’ by clinicians [113], the methods and/or instruments used, young maternal age, small sample sizes (especially in early studies and/or when analysing the effects of first and second degree unions) and the clinical conditions of the hospitals and/or nutritional/infectious problems in the place of birth, thus a downward trend is expected with better designed studies. After controlling for several of these non-genetic variables, Grant and Bittles still reported increased odds of post-neonatal (odds ratio, OR = 1.28), infant (OR = 1.32) and neonatal mortality (OR = 1.38) for the progeny of first cousins in Pakistan [130]. Another study carried out on Pakistani migrants living in Norway, has associated consanguinity with a seven-fold increased risk of progressive encephalopathy [131].

Reduced fertility was expected in consanguineous couples in relation to outbred individuals due to relatively higher chance of sharing specific human leukocyte antigen (HLA) haplotypes* causing failure to initiate pregnancy [128], or due to increased probability of deleterious variants being in homozygous state in genes which are essential for early embryonic or foetal development, [128, 132]. However many studies have shown that the exact opposite maybe true with 0.08 additional births per family on average [133-136], suggesting a compensating mechanism [105, 132]. Possible reasons given for increased fertility in consanguineous marriages (especially first cousins) [137, 138] are, even though owing at least in part to younger maternal age at first live birth [107], the relatively higher genetic compatibility between the mother and developing foetus – reducing sterility rates as well as prenatal losses. Also after first pregnancy, subsequent birth intervals are relatively shorter in consanguineous couples in part due to the lower likelihood of use of reliable contraceptive methods in consanguineous individuals [107]. All these can result in the optimisation of the maternal reproductive span [133]. Adding to the problem of confounding, previous studies have associated high consanguinity with low socioeconomic status (based on many factors such as occupation, household density, expenditure on food, education, lifestyle), illiteracy and rural residency which all have a direct influence on infant and early childhood mortality and morbidity, and therefore complicates dissecting the effects of consanguinity *per se* on mortality and morbidity rates in these populations [86].

Controlling for population stratification (PS)[†] is also be essential as crude comparisons between offspring of consanguineous individuals in relation to progeny of unrelated individuals will yield biased results in the presence of PS - as population specific mutations can confound estimates of morbidity and this will overinflate the effects of consanguinity *per se*. This was exactly the case, as in many

* It is a combination of alleles at adjacent loci on a chromosome that are inherited together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci

† It is the presence of a systematic difference in allele frequencies between sub-populations within a sample (most probably due to different ancestry). This is especially important in the context of association studies

other previous studies [113], in a 5 year prospective study which reported a 3 fold increase of post neonatal mortality and childhood morbidity in the offspring of consanguineous individuals of Pakistani background living in the UK [139]. They did not account for PS which is a known feature of indigenous and migrant Pakistani populations [140], thus the effect of consanguinity *per se* has been overestimated. Comparisons must be made within a homogenous population such as individuals from the same clan and/or tribe [141]. Failure to control for interactions between consanguinity and social variables (e.g. quality of life and health services, maternal age, maternal education, birth interval, birth order, demographics) would inflate the biological effects of consanguinity, leading to biased results. These factors make the previous literature all the more unreliable; and therefore, if care is not taken, contradicting results become inevitable. A few studies have not been able to replicate the increase in postnatal mortality (i.e. deaths immediately after birth) or rates of congenital heart disorders in the offspring of first cousins [142-144], as the excess of 1.7% to 2.8% reported is mostly attributed to autosomal recessive mutations within these populations which are in a homozygous state [105, 113]. The same is likely to be true with foetal wastage (i.e. foetus death in uterus), as no significant differences were observed between the offspring of consanguineous and non-consanguineous unions [145, 146]. Although a meta-analysis of miscarriages (i.e. loss of foetus before 20th week of pregnancy) showed an excess of 1.5%, significant outliers were shown to cause this [136]. Further complications arise if sample sizes are small or are focused on a single population as consanguineous marriages are practiced not only by the families belonging to lowest of socioeconomic strata, but also the highest [108]. Also within sub-structured populations, unrelated individuals have a higher chance of inheriting the same recessive mutation compared to ones living in large outbreeding populations. Therefore the health benefits of avoiding consanguinity, if any, is going to be less evident than what is expected by standard calculations [147].

Many studies fail to even indicate the different types of consanguineous unions (e.g. first cousin, double first cousin, uncle-niece) that resides within their category of

'consanguineous'. This can be another source of bias given the wide range of F values (range = 0.125 - 0.0156, and maybe even higher – discussed in section 1.3.4) within the definition of consanguineous unions, bearing in mind second cousin offspring are closer to non-consanguineous offspring ($F \approx 0$) than to offspring of first cousins ($F = 0.0625$); even more so to offspring of double first cousins and uncle-niece unions ($F = 0.125$). Therefore dissecting the effect of consanguinity solely on health outcomes requires careful consideration of covariates and confounding (see Figure 1.10 for a range of factors influenced by consanguinity and/or by living in regions where consanguinity happens to be prevalent), not just relying on a simple model of consanguineous versus non-consanguineous comparison [113]. These studies resemble ecological studies carried out in the field of Epidemiology which provide the lowest level of evidence* with many pitfalls (e.g. ecological fallacy).

Theoretically, consanguinity should not increase the risk of autosomal dominant conditions in offspring if one of the parents is affected; nor for X-linked recessive conditions if neither parent is affected [125, 126]. However, the Latin American Collaborative Study of Congenital Malformation (a large study based on over 34000 newborns) found a significant association between consanguinity and postaxial polydactyly (also hydrocephalus and bilateral cleft lip +/- cleft palate), which is found to be caused by autosomal dominant mutations (with incomplete penetrance). Even though they have controlled for confounders such as maternal age, ethnicity, maternal and paternal educational level, and paternal occupational level, residual confounding and/or differences in (disease causal) allele frequencies between the cases and controls could explain the findings. Other studies have linked consanguinity with aortic anomalies, non-syndromic neural tube defects, ventricular and atrial septal defects, pulmonary atresia, tetralogy of Fallot and patent ductus arteriosus [129, 148, 149], but later studies have not been able to replicate these findings indicating [105], just as aforementioned, population specific founder mutations [86, 129] (also see references [150-153] for lack of replicated negative and positive associations of consanguinity with cardiovascular disease and some types of

* Highest level of evidence is provided by randomised control trials (RCT)

cancers). Many studies have also not adjusted for multiple testing, with many studies reporting P values near 0.05 as ‘proof’ indicating the lack of statistical ‘know-how’ within the field.

A study demonstrating the critical and ubiquitous role of rare autosomal recessive alleles in an isolated and therefore inbreeding/endogamous population was carried out in the Dalmatian Islands of Croatia. The three papers published from these long term studies associated inbreeding with a variety of common disorders such as osteoporosis, hypertension and coronary heart disease [154-156]. However, as the authors also suggested, these findings are most probably all caused due to the recessive founder mutations which are found in the population and not due to inbreeding *per se*. Also the definitions used in the study are quite imprecise as ‘inbreeding’ (in the form of island endogamy) and ‘consanguinity’ are interchangeably used without clear separation [86, 156]. However studies such as these provide essential information for population and evolutionary genetics, and the implications of their results are discussed in section 8.1.

Thus so far, there is some but no *conclusive* evidence* connecting consanguinity with increasing risk of chromosome copy number variations/disorders (e.g. in the case of Down syndrome [105]), traits such as stature or with any complex disorder (e.g. intelligence quotient, cardiovascular disorders, schizophrenia, cancer, diabetes) to date [108, 113]; and quantifying any existing risk, if at all, will require further large-scale studies. Further studies on the genetic mechanism behind common complex disorders (as discussed in section 1.2.6), is also going to shape our way of understanding the effects of consanguinity *per se*. In theory, consanguinity would be expected to have a greater influence on a complex disease if the rare variant model is to explain the aetiology of a common complex disorder; and a variable, but mostly a lesser effect would be expected in the broad sense heritability model dependent on the influence of environmental factors on the disorder analysed. The infinitesimal

* None of the studies are backed up by studies where large samples sizes have been reached with replication in an independent cohort, and confounders such as population stratification and socio-economic variables have been added into the statistical model; or confirmed by functional studies

model would predict consanguinity's effect on common disorders to be of a very small effect brought about only due to higher levels of homozygosity in offspring of consanguineous unions (e.g. effect of minor allele is doubled in homozygotes compared to heterozygotes).

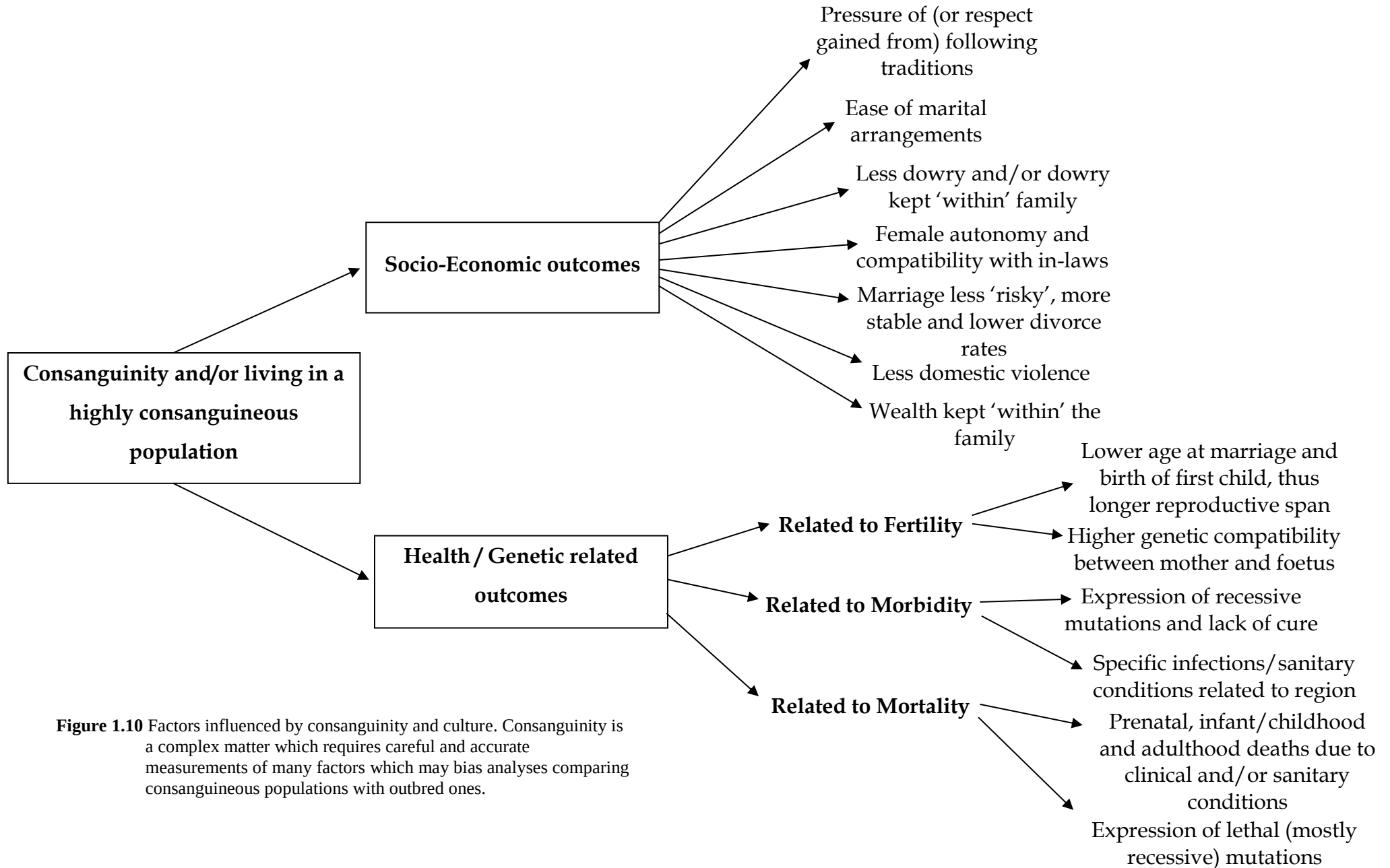


Figure 1.10 Factors influenced by consanguinity and culture. Consanguinity is a complex matter which requires careful and accurate measurements of many factors which may bias analyses comparing consanguineous populations with outbred ones.

1.3.3. Historical perspective

The English word ‘consanguineous’ originated in the early 17th century from the Latin word ‘consanguineus’ meaning ‘of the same blood’* – which is derived from the words ‘con’ meaning ‘together’ and ‘sanguis’ meaning ‘blood’ [157].

Contrary to current public belief in Western countries, consanguineous unions have not always been viewed as the source of “*appalling amount of defect and degeneracy*” [158]. They have been practiced amongst earlier societies; even amongst family members with high socio-economic statuses such as within royal dynasties (e.g. Egyptian Pharaohs, Holy Roman Emperors, Spanish Habsburg dynasty) and the ruling and/or land-owning elite to keep the wealth, “royal blood” and/or education within the family [106, 159, 160]. However, taboos have existed throughout history. Historically speaking, many scientists have blamed consanguinity as the main cause of inherited deficiencies. One even quoted:

“intermarriage has chiefly caused weakness of character leading to drink, not lack of brains or a certain amount of physical strength, but a very inert and lazy disposition” [161],

blaming consanguinity/endogamy for people’s laziness† and drinking of excessive alcohol in the island of Bermuda. This was because many at the time thought that the traits (including both physical and behavioural) of the parents were joined in the children in an additive way (i.e. ‘single dose’ of traits that the parents have in common results in ‘double dose’ in children) [158, 162]. First cousin marriage was certainly on the favoured side in Europe and North America up until the mid-19th century, especially amongst the elite [86]. Even Darwin married his first cousin Emma (Wedgwood) in 1839. Thus it is puzzling for some as to how consanguineous unions have become such an infrequent event so soon in these populations.

* Biologically speaking, the term is not the most appropriate as inherited characteristics are not passed on to the descendants via blood but through gametes which contain the genome.

† How laziness has been observed and quantified is another matter!

However the crucial role public debates amongst scientists of the mid-19th century played cannot be undermined as by the end of the century marriage between cousins was banned in 12 states in the USA [86, 163].

Negativity towards inbreeding is not hard to find in the recent historical literature either. For example some have described inbreeding as the major factor responsible for the extinction of the Spanish Habsburg dynasty (one of the two main branches of the 'House of Austria' alongside the Austrian Habsburg dynasty [164]) when King Charles II died in 1700 with no children. This claim is partly backed by genetic evidence (see reference [159]), although arguably it is far from being the major cause (e.g. small family sizes left very few heirs to the throne, decline in resources within the empire, health conditions of the era - as the infant and child mortality rate was very high). The written records reveal that formal disapproval has existed at least since the 6th century, the time of Pope Gregory I. He stated that:

"first cousin unions did not result in children and if they did, their offspring would not live a happy life" [86].

First degree consanguineous unions (i.e. union between siblings or parent and child) were frequent amongst the ruling class and/or dynasties of ancient Egyptian (i.e. Pharaohs) and Hawaiian societies, as well as in Zoroastrian Iran and the Inca Empire [104, 165]. Egyptian kings (and even some commoners) of the Pharaonic period (3200 BC - 332 BC) are known to marry their sisters/half-sisters as well as their daughters. This practice was also observed in the Ptolemaic period (305 BC - 30 BC) [165, 166]. As abovementioned, a more recent example (i.e. during the 16th and 17th century) is the Spanish Habsburg dynasty, with many kings practicing consanguineous marriages, including uncle-niece marriages as well as the usual first cousin unions [159].

Taboos towards incest (i.e. first degree unions) and consanguinity has existed in many human societies. For this reason there are many rules/laws that have been issued to prevent sexual relations between certain kin. These taboos have sometimes

extended to such an extent that even unions between a male and female with the same surname was prohibited in traditional Chinese societies*. Many theories have been proposed to be behind the establishment of incest and consanguinity laws, most prominent being the observance of undesirable effects in the offspring of close kin unions [104]. Also some theories on the origin of consanguinity and even incest have been associated with keeping the family as a 'stable' unit socio-economically and psychologically, thus some may have preferred unions with close kin. It is also undeniable that social attitudes and public behaviour influence choice to engage in union of kin. For example, genetically speaking, great-grandparents and great-grandchildren (which is a form of lineal kinship, $F = 0.0625$) are as related as first cousins (which is a form of collateral kinship) but the former type of unions are almost never observed due to sociological (e.g. views of the society and other family members) as well as biological factors such as age, physical attraction and fertility.

In relation to Pope Gregory I's comment above, it is fair to say religions have also played a part in shaping the history and prevalence of consanguineous marriages in many parts of the world. However there is a lack of uniformity even within the major religions. Consanguineous unions are prohibited in Orthodox Christianity. Unions between first cousins and farther related kin are allowed by the Roman Catholic Church and the Protestant denominations, albeit the former requires Diocesan permission [133]. Islam allows the marriage of cousins but contrary to belief there is no encouragement of consanguinity; and although the Prophet (s.a.w) married his daughter Fatima (r.a.) to his own cousin Ali (r.a.)†, several hadith (i.e. sayings of the Prophet) and sayings of the second Caliph Umar ibn Al-Khattab (r.a.)‡ have been found to endorse non-consanguineous unions. The Holy Qur'an has a comprehensive list of 'mahrams' (i.e. unmarriageable kin with whom sexual intercourse is considered incestuous – and therefore prohibited); and the list is similar to what is set out in the Old Testament (*Leviticus* 18) except that uncle-niece

* For this reason, a man could marry his mother's brother's daughter, but not his father's brother's daughter

† This would be an example of a union between first cousins once removed

‡ Umar (r.a.) has reputedly advised the Bani Assayib tribe to not marry cousins (see reference ¹⁶³)

marriages are not permitted in Islam. Religion is therefore the main reason why this type of consanguineous marriage, although highly consanguineous generally in relation to Western populations, are absent among the Arab population [113, 133]. Hinduism has been divided into two schools of thought in terms of marriage of kin. Aryan (Indo-European) Hindus strongly oppose marriage of biological kin with prohibition of unions for approximately seven generations on the male branch and five generations on the female branch [133]. In contrast, Dravidian Hindus strongly favour marriage within the family, especially between cross first cousins (e.g. mother's brother's daughter). In some parts of India, especially in the states of Andhra Pradesh (South Eastern India), Karnataka (South Western India) and Tamil Nadu (Southern India), uncle-niece marriages are also common (more than 20% of total Hindu marriages in Southern India [167]). Consanguineous unions are not allowed in Sikhism; however minor denominations within Sikhism do take sociological factors into consideration and allow them under certain circumstances. Buddhism in general, allows the couples and families to decide. Table 1.1 summarises the general attitudes of the main religions towards consanguineous unions. An academic reference for Baha'ism could not be found in the literature, thus a question was submitted to the National Spiritual Assembly of the Bahá'ís of the United Kingdom (NSABUK) to ask Baha'ism's view on consanguinity from a respectable source (see Figure 10.3 in section 9.2 for letter received from NSABUK)*.

Recent analysis of LD patterns amongst different World populations has backed up predictions stating that the founding population size of today's human populations had to be small [168]. With this in mind, consanguinity (and most probably incest) was inevitable [113], and almost certainly involved multiple occurrences of unions between close kin.

* I thank them for their detailed reply

Religion	General attitudes/rules on consanguineous marriages	Notes
Baha'ism	First-cousin marriage is allowed, although more distant blood relationships is advised	See section 9.2 for details.
Buddhism	Does not prohibit consanguineous marriages [169]	
Confucianism and Taoism	Consanguineous unions are permitted [170]	
Hinduism	Very different views between Indo-European (not allowed) and Dravidian (allowed) Hindus. [170]	The latter promotes consanguineous unions
Islam	Uncle-niece and Aunt-nephew marriages are not allowed but first-cousin unions are [170]	Both Sunni and Shias agree
Judaism	Consanguineous unions are allowed, including uncle-niece unions [170]	Aunt-nephew and half-sib unions are prohibited
Orthodox Christianity	Coptic: First-cousin unions are allowed [170] Greek and Russian: Unions between first-cousins and closer are not allowed [170]	
Protestant Christianity	Virtually all denominations allow first-cousin marriages [170]	Although culturally unacceptable in many societies
Rastafarianism	No references to consanguineous marriages found	
Roman Catholic Christianity	First-cousin marriages requires Diocesan approval [170]	Very strict rules apply[171]
Shintoism	Does not prohibit consanguineous marriages [169]	Mostly bound by Japanese law
Sikhism	First-cousin marriages are not allowed [170]	
Zoroastrianism	First-cousin marriages are allowed and widely practiced [170]	Religious endogamy is obligatory

Table 1.1 Views of main religions towards consanguineous marriages. NB: Where first-cousin marriages are allowed, lower levels of consanguinity are also allowed.

1.3.4. Autozygosity

So what has made consanguineous unions culturally unacceptable in many Western countries today? It is partly due to historical events (as discussed in section 1.3.3), the public not being informed well enough about the subject and partly because there is some truth behind having an offspring with a close relative being 'risky'. Consanguinity is not the 'great evil' it is made out to be [113] because there are many consanguineous families around the world (see section 1.3.5) without any affected members; in fact a large majority of them do not. The health risks arise from genetic factors which comes from the fact that union of kin results in an increase* in homozygosity (autozygosity†) in the offspring which elevates the probability of being homozygous for any allele which is within the family gene pool. If any of these alleles are recessive mutations, then this increase in probability also applies to these alleles. This simple fact is the main difference between consanguineous unions and the union of unrelated individuals; as there are many recessive mutations out there in every population and in every one of us, but because they are very rare (and probably unique to the families/individuals), they do not get to meet their counterpart in the union of unrelated individuals thus their homozygous effects are not observed (i.e. subjects remain unaffected). This is why rare autosomal recessive disorders are mostly seen in consanguineous families and/or regions where consanguinity (and/or endogamy) levels are high.

Identification of autozygous regions‡ and the quantification of total autozygosity (i.e. measuring the inbreeding coefficient for each subject analysed) represent a useful piece of information for genetic epidemiologists (Figure 1.11). This is due to the fact that the coefficient of inbreeding (also known as Wright's inbreeding coefficient [172], symbolised with an italicised 'F' hereafter) is used to define the probability that any allele be in a homozygous state and be derived from a common ancestor

* How much will depend on type of consanguineous union and level of inbreeding

† Homozygosity arising due to identity by descent i.e. sharing a recent common ancestor

‡ Also called 'autozygome'

(i.e. identical copies). For example, the coefficient of inbreeding in the offspring of first cousins would be $1/16$ (i.e. $F = 0.0625$) which means their progeny will be autozygous at 6.25% of genetic loci and therefore there is one in sixteen chance of inheriting a pair of alleles which are identical by descent (IBD)*.

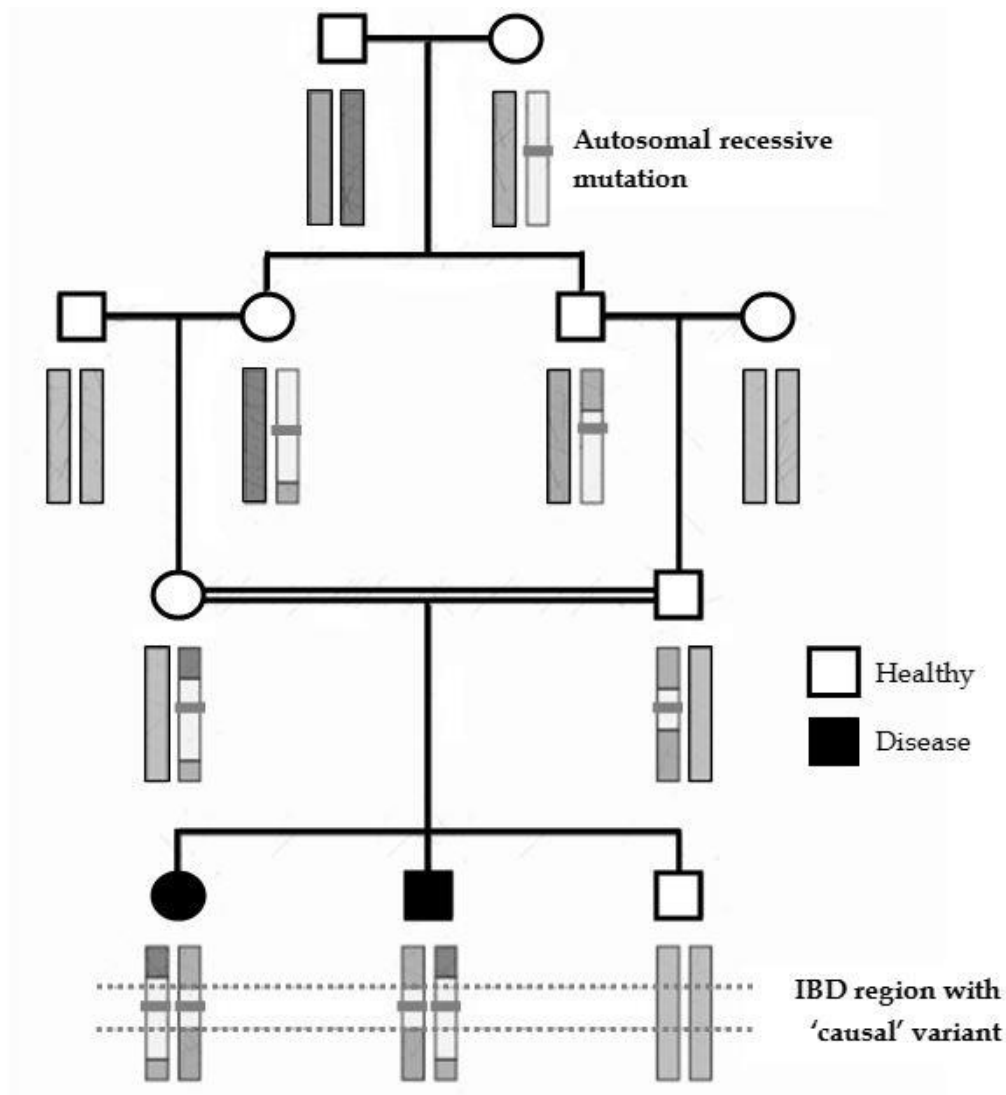


Figure 1.11 Homozygosity of two identical by descent (IBD) alleles. Autozygosity mapping has proven to be an effective way of pinpointing where a causal variant is located. With familial genotype data, autozygosity mapping becomes a relatively easier task, just as depicted in the figure above – by identifying the regions which are IBD. These regions are bound to get smaller through meiotic recombination as they pass down through the generations. However, sometimes even a single affected individual may be enough to identify the disease interval through determining of the IBD region with causal variant.

* Identical alleles which were inherited from a common ancestor

In clinical/medical genetics, any union between individuals who are related as second cousins or closer is considered 'consanguineous' ($F \geq 0.0156$) [173-175], thus the term includes first cousins once removed and double second cousins (see section 9.7 in appendices for depictions of all types of consanguineous unions). The threshold value has genetic implications as unions between individuals who are less related than second cousins are not expected (nor observed to) differ significantly from non-related unions.

The Wahlund effect* predicts higher levels of homozygosity in subdivided populations [176], therefore very high values of F ($F > 0.125$) are expected in 'isolates', small populations (and/or large families) whose members have been intra-marrying each other for many generations (e.g. consanguineous and/or endogamous). For example, Alvarez and Ceballos calculated the F value of (Spanish Habsburg dynasty) King Charles II (son of uncle and niece, Philip IV and Mariana of Austria) and Holy Roman Emperor Leopold I as 0.2538 and 0.1568 respectively [106]. Several other members of the Spanish Habsburg family also had F values higher than 0.20 (e.g. Marie Antoine of Habsburg with $F = 0.3053$ [106], daughter of Emperor Leopold I and his niece Margaret of Spain) [106, 159]. Note that the inbreeding coefficient of the former is higher than the expected F value (i.e. $F = 0.25$) for offspring of sibling unions (or parent-child). The standard Wright's inbreeding coefficients are calculated assuming that the grandparents are unrelated. This is not the case in isolated populations such as the Amish, Samaritans and Hutterites, many ethno-religious groups in North America; and the Druze in Israel, another ethno-religious group [148]. Also marriage within the family (most commonly between first cousins) is and has been strongly favoured by many populations in the Middle East and North and sub-Saharan Africa, as well as in Central and South Asia [104]. Pedigrees with complex loops arising from consanguineous marriages in successive generations are also common in these regions, which can also (and is shown to in [140]), bring very high F values to the fore. On the contrary, in outbreeding

* It is the reduction of heterozygosity in a population caused by subpopulation structure. An example of an underlying cause can be a geographic barrier to gene flow followed by genetic drift in the subpopulations

populations unions amongst individuals who are less related than first cousins (e.g. union between second cousins) can harbour F values which are close to that of a non-related couple ($F \approx 0$) [177]. Once the inbreeding loop is broken by the union of a consanguineous family member with an unrelated individual, the risk of autosomal recessive disorders in the offspring is expected to be lowered back to population background levels [147].

Solely using pedigree information to estimate levels of homozygosity has several limitations. As aforementioned, the standard measures assume that the grand-est parents (e.g. grandparents at the top of the family pedigree data obtained) are unrelated, thus does not take into account close-kin marriages that have occurred in distant generations, causing underestimation of the F values. Dependent on availability (and reliability) of data, standard measures also cannot fully integrate complex loops of inbreeding that may occur, especially in regions where consanguinity has deep roots (Figure 1.12). That is why modern approaches coupled with high density and genome-wide genotyping data are required to avoid the pitfalls of standard autozygosity measures/calculations. Analysing the genotype data obtained from the offspring of consanguineous unions for uninterrupted runs of homozygous regions (i.e. long runs of homozygosity, LRoH) has proven to be a reliable estimation method for quantifying total autozygosity (i.e. F value) of any individual. Studies carried out on European (and later in American) populations where estimates of autozygosity using length thresholds of 0.5Mb showed strong correlation with F values calculated from standard methods ($r = 0.86$), and demonstrated the applicability of the method [178, 179].

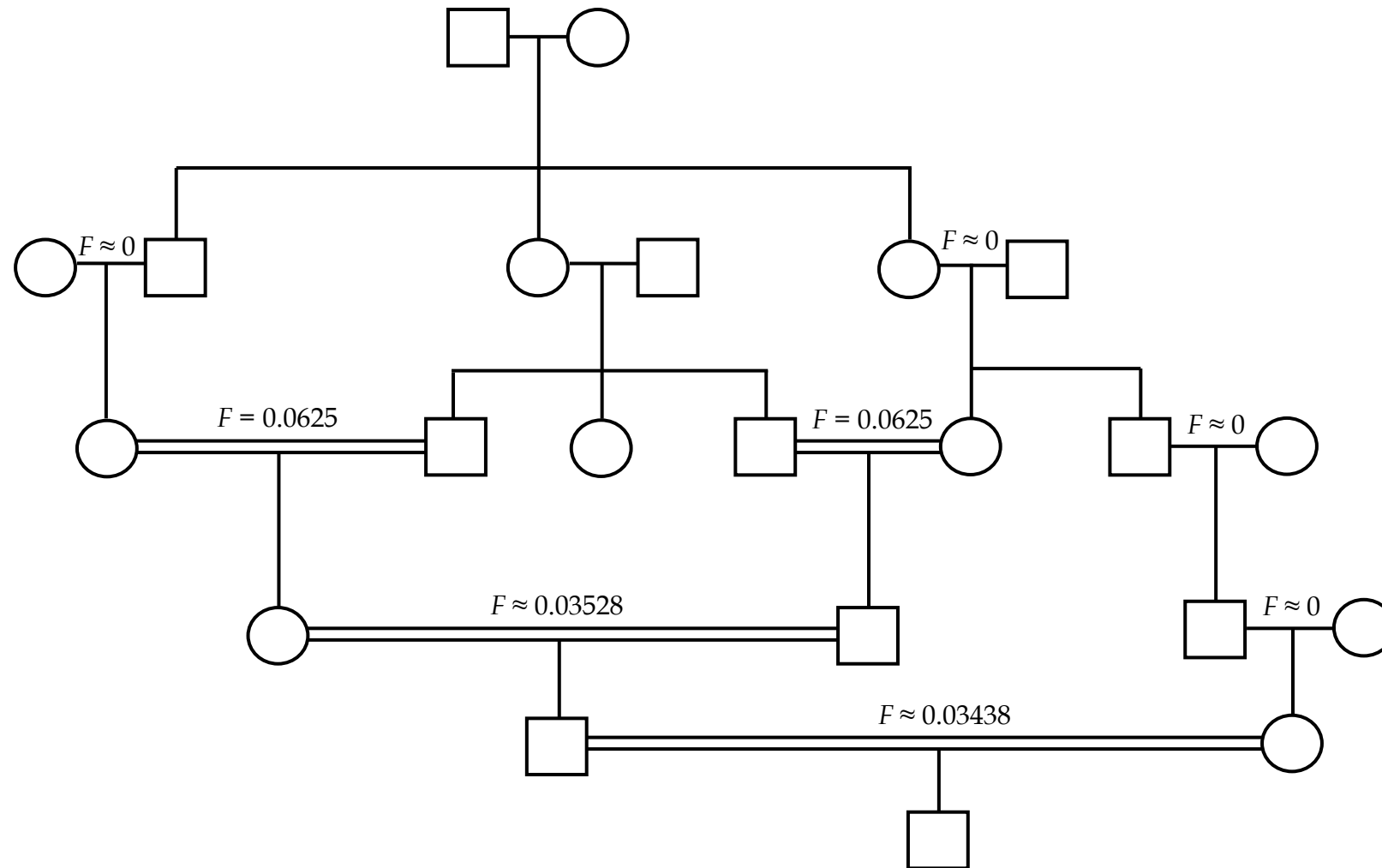


Figure 1.12 Example of a complex pedigree with multiple consanguineous unions. These types of complex loops of intra-familial unions which persist for many generations can harbour very large F values (especially if multiple uncle-niece and double first cousin unions occur). For example, even though the last consanguineous union is between second cousins ($F_{\text{expected}} = 0.015625$), the F value for their offspring is 0.03438 – which translates to approximately 120% higher autozygosity compared to standard second cousin offspring. This inflation in the F values could also occur in endogamous populations and/or tribes. Square: Male, Circle: Female, Double lines: Consanguineous unions.

1.3.5. World-wide Consanguinity

Consanguineous unions occur very rarely in Western countries for a variety of sociological (e.g. cultural, media coverage) and statistical reasons (e.g. smaller families means fewer cousins at similar age). However the complete opposite is true in certain regions of the world where union of kin is seen as the default choice, again due to many socio-economic reasons which will not be discussed in this section (see section 8.1 for additional information and discussion, but no great detail as it eludes the scope and objectives of this work). Analysis of currently available data suggests that over one-tenth of the World's current population is composed of families where the parents are related as second cousins or closer [86]. The prevalence of consanguineous unions was predicted to rapidly decline all around the globe in pre-World War II predictions; but not only have these predictions been proven to be false in most parts, the prevalence of consanguinity seems to have increased (except in Japan [180]) [133]. In fact it is thought to be increasing still in a number of countries such as Morocco [113], Qatar [181], the United Arab Emirates [182] and Yemen [183]. This is mostly attributed to greater number of children surviving to marriageable age, which enables social preferences in marriage (i.e. consanguineous unions) to be more readily accommodated [133].

Consanguinity is a highly respected social trend with deep roots spanning vast regions around the world (Figure 1.13) [105]. These regions are predicted to encapsulate one-fifth of the World's population (i.e. over a billion people) with most residing in the Middle East, West Asia, and North Africa where intra-familial unions account for twenty to fifty percent of all marriages [113, 133]. Also many migrant communities living in North America, Europe and Australia engage in consanguineous and/or endogamous unions [86].

The prevalence of consanguinity (especially first cousin marriages) is influenced by ethnicity, religion, culture and geography [105] – and thus, the vast difference amongst different populations. These factors are also shown to play a similar role

where immigrant minorities from highly consanguineous regions are clustered in other countries/regions where consanguinity is not common practice such as in Europe, North America and Australia [108, 137]. Many studies show that prevalence of consanguinity is on the rise in Western countries due to these migrant communities [125, 127, 172]; and the effect of sociological factors which incline members of these communities towards consanguinity in the first place are exacerbated (i.e. become more influential) after migration to Western countries (discussed in section 8.1). Another factor influencing consanguinity levels is law (Figure 1.14). First cousin marriages are either 'illegal' (thirty as of Nov 2014, counted as 'criminal offense' in five of them) or 'banned with exceptions' in most states in the USA [184]. Currently (as of Nov 2014), marriage between cousins is also prohibited in the Philippines, Bulgaria, Croatia and Romania. Also the Hindu Marriage Act of 1955 banned uncle-niece marriages in India, however it was deemed 'impracticable' especially in Southern India, and was later revoked in 1984 [86].

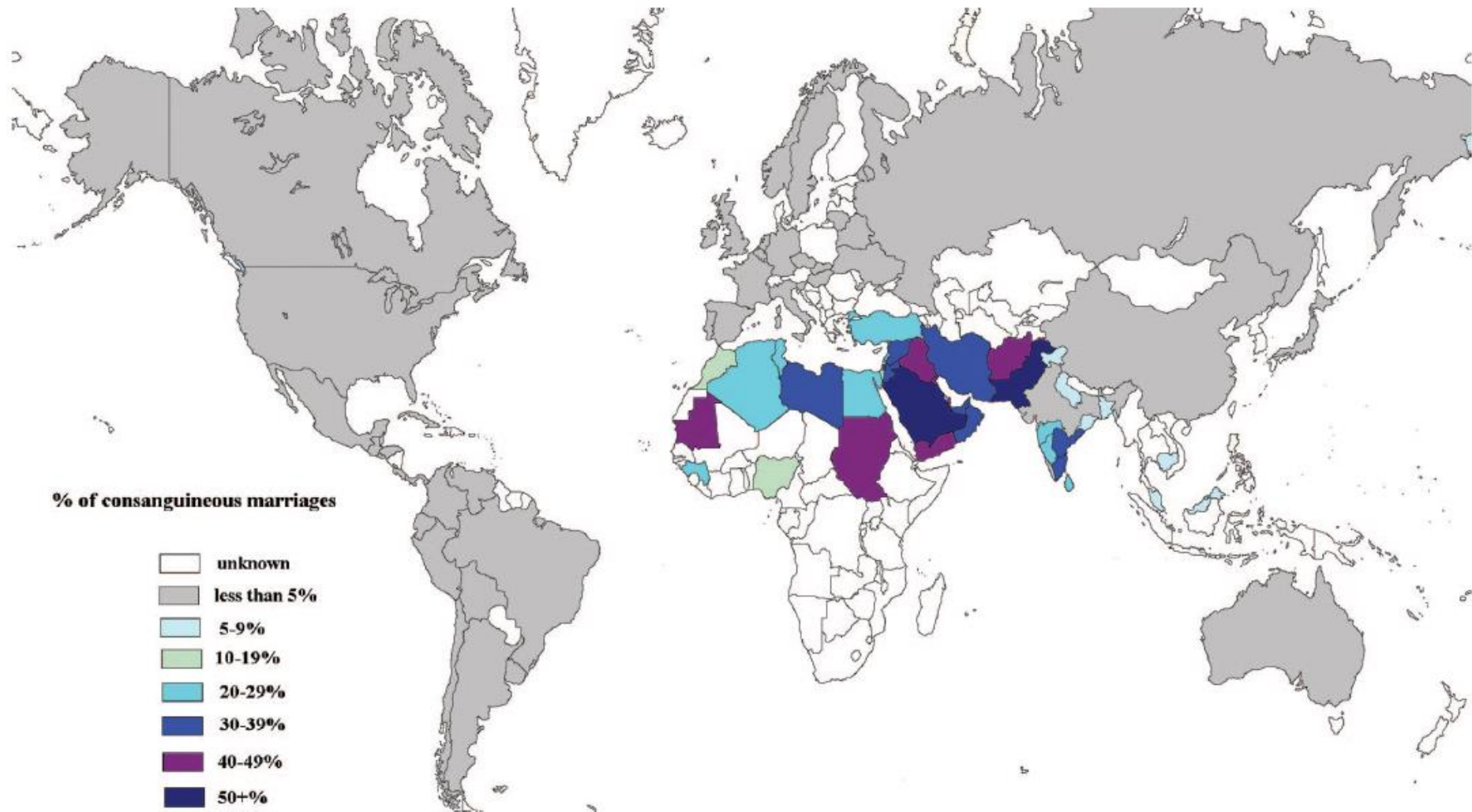


Figure 1.13 Worldwide consanguinity. Contrary to public view in the West, consanguinity is traditional and respected by many communities. Intra-familial unions collectively account for 20 to 50% of all marriages in the transverse belt that runs from Pakistan in the East all the way to Morocco in North-West of Africa [105]. As of current, there is no reliable data on consanguinity rates for most parts of sub-Saharan Africa and Central Asia. Figure reproduced with permission from Nature Publishing Group [105].

Nations can be arbitrarily divided into four categories in terms of consanguinity levels: (i) low - where consanguineous marriages account for less than 5% (ii) intermediate - between 5% and twenty percent (iii) highly consanguineous (over 20% and below 40%); and (iv) extremely consanguineous (over 40%). Also there are many countries, especially in sub-Saharan Africa, where consanguinity levels are unknown (or not reliably known) due to lack of extensive research. Conservative estimates predict that approximately one-sixth of the world's population (a figure of 1.1 billion is proposed by the Geneva International Consanguinity Workshop Report [105]) live in regions which fall into category (iii) and (iv) [133]; and also another one-sixth falls into the 'unknown' category – the latter fact reflecting the need for further research of consanguinity in all world populations, especially Africa where tribal connections are known to be important.

Highest levels of consanguinity have been observed amongst the families of Armed Forces personnel in Pakistan and in urban Pondicherry where the figures reach 77.1% and 54.9% respectively, with 20.2% of total marriages in the latter being uncle-niece unions [159, 185].

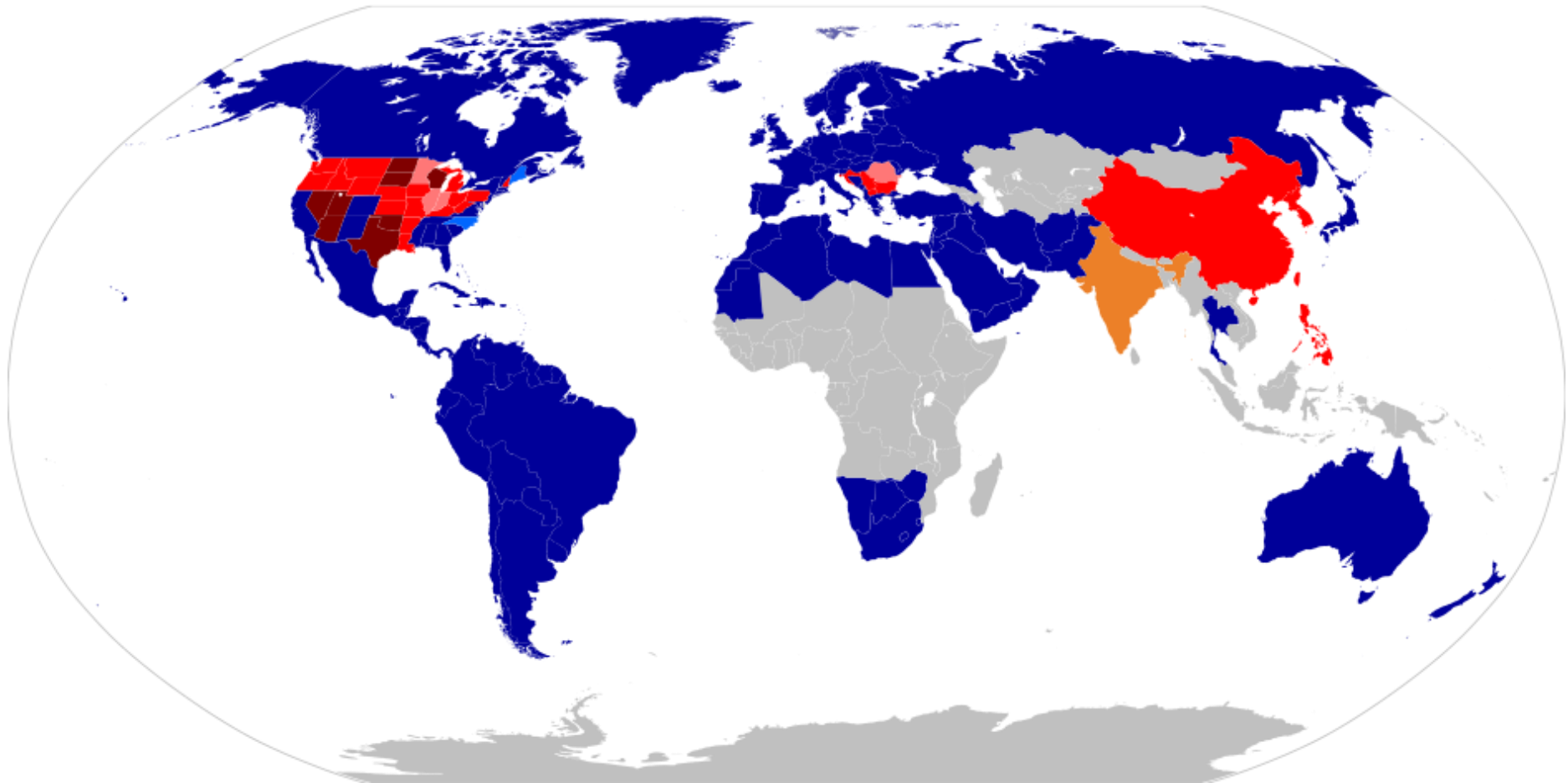


Figure 1.14 Laws regarding first-cousin marriage around the world. ■ First-cousin marriage legal ■ Allowed with restrictions or exceptions ■ Legality dependent on religion or culture ■ Statute bans first-cousin marriage ■ Banned with exceptions ■ Criminal offense ■ No available data. The image has been released into the public domain by the author (URL: http://en.wikipedia.org/wiki/Cousin_marriage).

1.4. Identifying the Genetic basis of human diseases

Many human diseases/disorders have a heritable component. Identifying the genetic basis of human diseases requires detecting *and* dissecting the causal variation(s) from the tens of thousands of neutral ones in the affected individual's genome [57, 186]. In an ideal world, a perfect DNA sequence of sampled individuals would be available without any need for post sequencing quality control measures; and by comparing affected and unaffected individuals one would be able to determine which variant(s) has been responsible for the disease or trait. However, even with the availability of recently developed NGS technologies, (completely) accurate sequencing and mapping of the reads to the reference genome (or parts of the genome) still represents a considerable problem. Not helping the cause – even with all the molecular and clinical tests available, accurate phenotyping can also be a nuisance for Mendelian and complex disorders which do not have distinct symptoms and require deep phenotyping (e.g. misclassification of some Primary ciliary dyskinesia patients is still a major problem). The former problem causes a considerable quantity of NGS sequencing data to be discarded as many variants fail one or more of the standard quality control filters; and the latter one (i.e. misclassification) can lead to wrong conclusions and “causal” variant to be reported, especially if the sample sizes are small (which is often the case in familial studies).

Once a candidate mutation has been identified, complementary functional and comparative studies should be carried out in model organisms* to backup causality. This is required as mutating the identified gene in another individual cannot be carried out for ethical reasons. Thus a homologue of the gene is searched for in a model organism.

* Where a homologue exists

1.4.1. Traditional methods

It is easy to forget that the human genome was mapped just over a decade ago. With the development of relatively cheap SNP arrays and DNA sequencing methods, sequencing (or genotyping throughout) the whole human genome (or exome) enabled genetic analyses to be (relatively*) hypothesis-free and enabled researchers to carry out unbiased genetic association studies (i.e. ascertainment bias in candidate gene studies). Before this, only low-resolution linkage analyses or a selection of candidates for association testing was possible – mainly for technological and cost reasons [187]. Candidate genes were determined from previous literature based mostly on studies of model organisms, thus it is inevitable that many candidate gene studies carried out in humans have remained inconclusive. Publication bias may be covering up the true extent of the time, resources and funding that has been spent on these studies [188].

Linkage analyses capitalised on the fact that recombination is less likely to occur between genetic markers which are in close proximity in relation to ones which are further apart or in separate chromosomes [189, 190]. If the probability of segregation of the markers is less than 50%, then they are said to be linked (i.e. on the same chromosome), which results in the co-segregation of the markers more often [191] than if they were unlinked (i.e. independently inherited). The pre-genomic era only enabled researchers to use a relatively low number of markers spread across the genome. Thus analysing which markers were co-inherited with the disease facilitated only a low resolution mapping of where the disease causal mutation lies. Also this approach was successful only if there was a large pedigree with multiple affected individuals[†] and worked mainly for truly monogenic Mendelian disorders[‡].

* The word 'relatively' is used because one is still assuming there is a genetic component to the disorder or trait being analysed

[†] A LOD score of 3 (Prob= 0.001) had to be reached

[‡] As genetic heterogeneity causes loss of statistical power

1.4.2. Current methods

Mapping the human genome (for the first time) coupled with technological advances (e.g. computational, DNA sequencing platforms) have paved the way for genetic disorders to be analysed in a more automated manner. These breakthroughs lowered the costs of analyses and caused a rapid increase in the amount of organisms which have had their whole genome sequenced and mapped (see GenBank growth stats at: <http://www.ncbi.nlm.nih.gov/genbank/statistics>). Also many projects mapping human variation between and within populations were initiated (see section 1.7). A few years after the completion of HGP, the first genome-wide association studies were carried out where millions of genetic markers (i.e. SNPs) across the genome (determined mostly by the HapMap project) were genotyped in large numbers of individuals; and consequently the difference in the frequency of these SNPs were then compared across different groups (e.g. cases and controls) in order to identify genes/regions that contribute to the disease or trait under analysis.

The theory of some of today's methods were thought of well before the millennium, however they were not practically possible due to costs and lack of automated platforms. One of those was that millions of markers distributed across the genome could be tested for association with any trait using cases and controls [192] – which is the idea behind today's GWASs. The initial GWASs were low powered and issues such as multiple testing*, cryptic relatedness† and population stratification were not very well understood. These problems were solved soon enough through experience and the establishment of large consortia [27, 193-196]; and today GWASs have become one of the most successful types of genetic association studies with findings in many GWASs replicated in follow up studies [194-201]. Results from GWASs are

* Setting $\alpha = 0.05$, one would expect one in every 20 tests to produce a false positive. Therefore the more independent tests carried out, the α should be decreased accordingly e.g. by using a Bonferroni correction

† Kinship between a few cases (or controls) that is not declared to the researchers – violating the 'independent data' assumption of linear regression used in GWASs

collected in the NHGRI GWAS catalogue [81] – which is publicly available at <http://www.ebi.ac.uk/gwas/>.

An alternative emerging method is exome-wide association studies (ExWAS), which makes use of WES technologies in contrast to the genotyping across the whole genome which occurs in the former. Sequencing data will enable association of the variants themselves rather than using a SNP to ‘tag’ an LD block*. Carrying out association studies using WGS data is the ultimate goal which may become a reality in the next decade. However the statistical, computational and bioinformatics needs of such projects have already got scientists thinking rigorously about potential pitfalls.

With the completion of large scale projects such as the HapMap, NHLBI GO Exome Sequencing Project (Exome variant server, EVS) and 1000 Genomes projects (1000GP, details in section 1.7.1), the vast amount of freely available genetic data and the development of comprehensive bioinformatics software and computers which can handle the enormous amounts of data produced, unearthing the genetic component of *any* disease is now feasible, albeit not yet perfect. This in return has and is going to enable the development of cheap but reliable diagnostic tools - and in long-term, new therapies.

The next section will discuss how DNA sequence variation is detected and how detecting these can elucidate the genetic basis of human diseases.

1.5. Detecting DNA sequence variation

Traditional cytogenetic techniques allowed clinicians to extract and visualise the chromosomes of affected individuals with dyes such as Giemsa (in G-banding†) and quinacrine (in Q-banding). Cytogeneticists would then manually analyse the affected

* A population based phenomenon whereby several variants in close proximity in the same chromosome are co-inherited (much) more often than expected by chance – hence the term ‘block’. LD blocks are population specific, thus a ‘tag’ SNP used in one population may not do a good job in another population; even worse, may be monomorphic.

† Heterochromatic regions stain darker with this technique

individuals' karyotypes to check whether the right number of chromosomes were present and whether every single band was present in the right size and region. However this technique only allowed whole chromosomal copy number variations, large scale (i.e. cytogenetic) deletions/insertions and translocations (e.g. 5 to 10 million base pairs or larger [202]) to be identified; and any variation which required higher resolutions were left undetected. A famous example of the era was the Philadelphia chromosome* which was associated with chronic myelogenous leukaemia [203]. Other approaches which make use of enzymes to detect restriction fragment length polymorphisms (RFLP) or use enhanced dyes to detect chromosomal aberrations with higher resolution using fluorescence *in situ* hybridisation (FISH) represented significant advances in the field of DNA sequence variation detection [204].

The discovery of heat stable DNA polymerase, the development of polymerase chain reaction (PCR) and subsequent improvements on the PCR methods were arguably one of the biggest breakthroughs in genetic engineering. These advancements in the field allowed any primer to be designed which could then be used for many different purposes including detection of human DNA variants (using techniques such as Single-strand conformation polymorphism [205] and Denaturing high-performance liquid chromatography [206]), amplification and insertion of whole genes (or any sequence) and/or testing the presence/absence of certain genes/alleles in model organisms/humans [207]. These techniques are and have been successfully used to facilitate our understanding of the human genome. However all the above mentioned techniques usually test a single or a few loci at a time; and thus could take us so far, and techniques which could detect single nucleotide variation in an automated and high throughput fashion were needed. This would allow for faster, cheaper and above all, reliable variant screening approaches to be made available. Although first generation DNA sequencing techniques were around, they were very expensive to use. Where the technology was

* Identified by an abnormally short chromosome 22 that is found in the hematopoietic cells of persons affected with chronic myelogenous leukaemia and lacks the major part of its long arm which has usually undergone translocation to chromosome 9

available, they were mostly used in candidate gene studies – reviewing a handful of genes.

1.5.1. Whole genome sequencing

Detecting all variants in the genome of an individual is the holy grail of genetic research. This will enable viewing the full spectrum of mutations an individual has, and this is important for identifying a continuum of variants with ‘no clinical effect’ to ones with ‘detrimental’ effects.

Many bioinformatics tools have been developed which tries to utilise the publicly available WGS data and impute sequences of many individuals who have had hundreds of thousands of SNPs genotyped using genome-wide SNP arrays. This is done by deriving haplotypes from a large set of individuals who have had their whole genome sequenced. One of the primary aims of the 1000 Genomes Project (1000GP) was exactly this: to whole sequence the genome of thousand individuals from distinct populations which would in return serve as reliable reference haplotypes which can then be used in other analyses such as defining LD blocks, performing genotype imputation* and haplotype phasing.

Imputation is useful as performing WGS on every single individual is still not feasible due to costs. However the main limitation of imputation algorithms is that it is unreliable in determining the genotypes of rare variants. Also WGS data obtained from one population may lead to bias in another population due to difference in allele and haplotype frequencies. The high costs of WGS is such a limiting factor that even state funded projects such as the 1000GP [208] and more recently the UK10K project opted to use sequencing at low read depths† in order to reach the sample size

* It is the statistical inference of unobserved genotypes through the use of known haplotypes in a population (e.g. from the HapMap or the 1000GP), thereby allowing to test initially untyped genetic variants for association with a trait of interest. Genotype imputation hence helps tremendously in narrowing-down the location of probably causal variants in GWAS

† The term refers to the average amount of sequence reads mapped onto a reference sequence

targets they had initially set. This also makes imputation less reliable. These limitations can only be overcome with performing deep WGS.

Until the development of next generation sequencing (NGS, discussed in section 1.6.2) technologies, sequencing the whole genome of a single individual was possible only through large international projects, but today, albeit costs still in the five digits sums, we can presume that the days where WGS will become common practice in most labs are very near. At current where sequencing is carried out, targeting the exome has become the pragmatic choice (see section 1.5.2). However WES will become obsolete and be replaced by WGS one day as the latter does not require the initial exome targeting procedures. This change to WGS will facilitate our understanding of the contribution of rare and/or noncoding variants on common complex disease greatly.

1.5.2. Whole exome sequencing

Due to the costs of WGS still being prohibitive for most, WES has been proposed as an alternative. By targeting only the regions within genes which code for proteins, classified as the 'exome' (which does not include the introns), the cost is pulled down approx. three fold in relation to WGS*. However although the exome represents less than 2% of the genome, WES is thought to capture around 85% of Mendelian disease causal mutations [59].

Efficiency and reliability of WES has improved considerably in relation to previous exome capture kits. The target enrichment process has been improved with the development of a variety of reliable methods such as array based [209, 210] and solution based hybridisation [211]. This improvement in performance is also down to the plethora of bioinformatics tools that been developed. Analysing only the exome decreases the amount of storage needed and the time and effort needed to

* WGS generally costs ~\$3000 (higher or lower depending on offers and sample size), whereas WES costs around \$1000. At present, GATC-Biotech can offer WES (at 60X) for as low as £500 per sample for large sample sizes (personal communication)

analyse in relation to WGS*. Currently, we best understand the coding part of the genome which makes WES also the pragmatic choice in terms of prioritising and analysing candidate genes and/or variants [210, 212-214]. As WES targets all exons†, it provides an unbiased view of variation across all known coding regions, which is another advantage over other targeted sequencing methods such as candidate gene sequencing studies (not WGS) and genome-wide SNP arrays.

Many researchers who analyse Mendelian disorders have embraced WES as it virtually provides all the advantages of performing a WGS but itself does not come with the sequencing/labour costs and storage/software needs of WGS. Of course there is an obvious risk of implementing WES which can lead to an unsuccessful analysis: which is when the causal variant(s) resides in a noncoding region (see example of Usher syndrome [215]). Other limitations of WES are brought about with the chosen exome capture kit, which will only attempt to capture their choice of (currently known) exons [186]. It is also reported that some regions with very high GC content are not captured efficiently as other regions [186]. Additionally, as with WGS, there are bound to be areas with low read depth and reads with sequencing errors brought about as machine artefacts [216, 217]. Finally, in concordance with the findings of the ENCODE project, exome capture kits miss out on a lot of functional elements which are shown to have a significant role in disease aetiology [218]. Albeit it is still a relatively new technology (commercial kits were made available in 2011), WES has been proven to be successful in identifying the causal variant of a variety of Mendelian disorders, even previously undiagnosed ones [82, 186, 217, 219-221].

After comparison of the query sequence with a reference sequence (e.g. hg19), approx. 20000 coding SNVs (this number will depend on many factors e.g. ancestry) and indels are identified per individual [82]; and pinpointing the causal 'one(s)' is the main challenge. Various filtering methods can be used to reduce the amount of non-causal candidates and this will depend on many factors such as the sample size,

* This is true even when compared to the Sanger sequencing and multiplex PCR methods ²³¹, considering the amount of exons sequences (~180 thousand)

† Plus several bases from either side – depending on the target enrichment kit used

nature of the data (e.g. consanguineous collections, unrelated individuals), nature of disorder analysed (e.g. genetic heterogeneity, mode of inheritance) and previous literature (e.g. candidate genes, known variants). Also the availability of population based data can be crucial to filter common variants, unlikely to be causal of a rare disease.

However as aforementioned, in the long term WES is likely to be replaced by WGS. As the price of sequencing decreases, the cost and time that is spent during the whole exome enrichment step will become unnecessary to carry out [213].

1.5.3. Other methods

An alternative approach to WGS and WES is to sequence certain regions/genes. This can be extremely effective (e.g. high read depth, less analysis, less storage needs) and will ultimately cost much less especially if the causal variant is known or suspected to be in the chosen region(s). This region can be pinpointed in consanguineous collections or in populations with a small number of founders firstly by dense genotyping (across the genome) and then identifying the overlapping long runs of homozygous (LROH) regions in affected individuals. Then the regions that are present in unaffected individuals and/or not present in all the affected individuals can be discarded leaving a very few number of, if not one, regions to then sequence.

Current sequencing techniques struggle to accurately pinpoint CNVs larger than 50 bases as the reads produced from the sequencing machines do not capture the whole CNV thus when the reads are mapped to a reference genome, the algorithms struggle to distinguish non-overlapping reads amongst the hundreds of similar reads. Array based comparative genomic hybridization (aCGH) is a popular and more accurate method in clinical genetics when analysing CNVs across the whole genome. DNA obtained from a control and a patient is sheared into smaller pieces (e.g. using enzymes, PCR, shearing) and labelled with different fluorescent dyes. They are then hybridised on a solid (e.g. glass, plastic) slide which contains

thousands of DNA capturing probes. Once the non-hybridized DNA segments are washed off, the hybridised DNA from the case-control duo will result in a distinct colour* and intensity; and this information can then be used to deduce the amount (and type: deletion or insertion in relation to the reference) of CNVs present in the patient [222]. Comparison of these methods in relation to different types of analyses is summarised in Table 1.2 below.

* For example if the control DNA fragments are coloured red and the affected individual's DNA green, after mixing, green areas will indicate insertions/duplications in the latter DNA; and red vice versa.

Technique/Technology	Large CNVs	Large-scale Structural changes (exc. BT)	Balanced translocations (BT)	Consanguinity and/or Endogamy	Uniparental disomy*	Novel (coding)	Novel (non-coding)
Targeted gene sequencing	N	N	N	N	N	Y	N
SNP arrays	N	Y	N	Y	Y	N	N
FISH	Y	N	Y	N	N	N	N
Array CGH	Y	Y	N	N	N	N	N
WES	N	Y/N	N	Y/N	Y/N	Y	N
WGS	Y/N	Y	Y	Y	Y	Y	Y

Table 1.2 Clinical potential of widely used methods. Y: Yes, detected, N: No, Y/N: Partially. Table adapted from Alsolami *et al* [223].

* Uniparental disomy occurs when a person receives both copies of a chromosome/region from one parent

1.6. DNA Sequencing technologies

Sequencing the whole genome became a reality after the discovery of dideoxynucleotides (ddNTPs) and their important use in the dideoxy chain terminator machines. Before sequencing, disease associated loci were picked up in low resolution by genetic markers but the exact causal variant and/or gene could not be determined. Technological advances both computationally and in laboratory-based settings (e.g. capillary electrophoresis, multiplexed PCR machines) led to the development of new methods and concomitantly decreased the costs of performing studies [224]. Today, Sanger sequencing machines are almost never used to sequence a genome or a large region of it; and have been replaced by 'next generation sequencing' platforms. This change, as aforementioned, has reduced the price of sequencing the whole human genome from billions of dollars to five digit sums (or four, depending on currency); and the time required for completion from years to days – and soon to be hours.

1.6.1. Historical background

The first complete gene sequence to be published was achieved by Fiers' group in 1972 using RNA sequencing [225]. Five years later, two novel DNA sequencing methods were published in the same year – first by Maxam and Gilbert [226], and then by Sanger's group [227]. Maxam and Gilbert's method, also called chemical sequencing, involved applying a procedure that breaks terminally labelled DNA molecules at each repetition of a base; which were then resolved by size via gel electrophoresis. Albeit accurate, the former method's use of radioactive labelling and high technical requirements made Sanger's method more popular – especially after further refinements were made on the latter method.

Sanger's first paper was published a few months later than Maxam and Gilbert's. In that seminal paper he explained how radioactively labelled ddNTPs can be used to

terminate the polymerisation of the next dNTP [227]. Further improvements on the original idea with the addition of fluorescent labelling, capillary gel electrophoresis and extra automation enabled (for that time period) relatively high throughput, cheap and accurate sequencing machines to be developed. The HGP was completed in 10 years using the (refined) Sanger method with the use of hundreds of machines running in parallel [228]. The Sanger method required labour intensive and time consuming DNA enrichment steps (via PCR) and was limited to just over 90kb in each run [229]; and the running costs of experiments were still too costly for a lot of laboratories which did not receive sufficient funding. Thus there was a need for new methods which would attempt to solve these problems. However, not many could have seen the technological advances and the enormous decrease in prices of sequencing that occurred within a matter of years after the HGP; the age of next generation sequencers was about to begin*.

1.6.2. Next-generation sequencing

The term next generation sequencing (NGS) is used to describe all sequencing platforms succeeding the traditional Sanger method described in the previous section. Thus NGS encapsulates 'second' and 'third' generation sequencers. The difference between the two is that the former requires amplification of target molecules prior to sequencing with huge numbers of reads produced [230, 231] whereas the latter sequences individual DNA molecules in real time and do not require the pre-sequencing phase amplification. These technologies have contributed to genetic research by reducing the cost of sequencing which in return increased the amount of sequencing projects carried out. These technologies are also called 'high throughput' as they allow the whole 3 billion base pairs of the human genome (or whatever is inputted, e.g. whole-exome) to be sequenced in a single run [232].

* There is a fantastic poster on the 'Evolution of sequencing technology' made available by the Science magazine. It is available online at: <http://www.sciencemag.org/site/products/posters/>

There are a variety of NGS technology companies competing with each other (e.g. Illumina with HiSeq 2000 and MiSeq platforms, Life Sciences with SOLiD and Ion Torrent PGM platforms, and Roche with the GS FLX Titanium and GS Junior platforms), which is an indicator of further decrease in sequencing platform and running costs. Each use a different method to carry out the sequencing process, thus each bring with them their own advantage and disadvantages. For example Illumina takes pride in providing low error rates and low cost per base making them the standard for resequencing projects, whereas Roche 454 outputs longer read lengths making them the standard for *de novo* sequencing of new genomes, especially large genomes where post-sequencing mapping of the reads is crucial. Comprehensive and comparative information about NGS technologies can be found in a expertly curated blog at <http://www.molecular ecologist.com/next-gen-table-4-2014/> or in a review article by Glenn (233).

1.7. Population-based genetic variation datasets: why collect them?

The HGP and subsequent large scale projects such as the 1000 Genomes Project have shown that there is a great deal of variation even in a single person's genome (relative to a reference genome), let alone the whole human population of over 7 billion (population data from [234]). Although understanding the effect of all possible variants is the ultimate challenge, this is a utopic idea for now. Thus the priority is on clinically important (e.g. disease causal) ones. A single genome has between three and five million variants (coding and noncoding, depending on ancestry) [235]. However out of these, less than 5000 of them are nonsynonymous variants (i.e. stop-gains, missense) and exonic indels – which are usually the main culprits in disease phenotypes [235].

1.7.1. Projects for mapping human genetic variation

Each individual is phenotypically different from one another; and DNA sequence variations at defined positions within the genome are responsible for a majority of these. These variations not only change the way we look on the outside, but also may increase one's propensity for a common complex disorder such as cancer or diabetes. To understand the variation that exists in the human genome across (and within) populations, many large-scale international projects were initiated such as the HapMap (in three phases) and 1000 Genomes projects. Also other individual laboratories and consortia were encouraged to share non-confidential* data publicly to advance genetic research – where informed consent was obtained. Many publicly available whole-genome and gene specific databases were made available with each serving a different purpose. Whilst sequencing the whole genome (or exome) of individuals was still prohibitively expensive, the HapMap project used the cost-effective alternative of genotyping millions of validated SNPs of 270 individuals from different four different populations (i.e. Yoruban, Han Chinese, European ancestry and Japanese) in order to identify marker SNPs which can be used as a proxy for other SNPs in the same LD block – reducing cost, which in return allows larger sample sizes and highly powered studies. More information on the HapMap project can be found at <http://hapmap.ncbi.nlm.nih.gov/>.

The 1000 Genomes Project differs from the HapMap project in that the former carried out WGS on over 1092 individuals instead of the dense SNP genotyping which was carried out in the latter. Again the motive was similar, which was to map genetic variation (including rare variation unlike the HapMap project which concentrated on SNPs with MAF over 1%), whilst also providing researchers with established reference haplotypes which can be used for imputation of ungenotyped loci in GWAS. WGS picks up rare variants, however large sample sizes are required

* Dense genetic data is always going to infringe confidentiality rules, as everyone has their unique (combination of) variants which will allow bad-intentioned people to trace it back to the donor. However by releasing the variation data summary statistics for the *whole* sample (e.g. MAF), datasets owners can alleviate these fears

for the reliable imputation of these variants as imputation accuracy is directly related to allele/haplotype frequency. The more WGS data that is available, the more accurate the imputation process is going to get*. All detected variants are available freely at dbSNP (and <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release> for download) and also can be added as annotations to VCF files via Ensembl VEP (<http://www.ensembl.org/info/docs/tools/vep/index.html>).

Another project worth mentioning is the NHLBI Exome Sequencing Project (ESP, current version ESP6500), where the whole-exomes of over 6500 unrelated individuals (4300 European-Americans + 2203 African-Americans) were sequenced with high read depth (>50x on average) in order to understand the contribution rare coding variants (i.e. SNVs and Indels) make to complex disease [214]. The ESP study identified more than half a million variants of which over 80% were classified as rare (with a MAF of <0.5%), which includes over 95% of SNVs predicted to be functionally important [214]. The exome data has been made available on the database of Genotypes and Phenotypes (dbGAP), dbSNP and Exome Variant Server (EVS), and the latter can be incorporated via Ensembl VEP.

Since then, in mid-October 2014, the ExAC database (beta) was made publicly available which contains the MAF data of 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC database can be found at <http://exac.broadinstitute.org/downloads>.

The abovementioned, large human genome variation projects have enabled the establishment of well-annotated databases. They facilitate the filtering of variants known to be benign or presumed to be non-causal[†] of the disease under inspection during the analysis stage. These databases can also be used in addition to internal databases many research councils, labs and/or universities have. Three of the well-

* Ethnically matching the query sequences (to be imputed) with population data is also crucial and projects such as the UK10K project (combining ALSPAC and Twins UK) are one of the few initiatives which is trying to offer a solution to the issue.

[†] For example they may be too common to be causal of a rare disease

known databases of genetic variation are dbSNP, the GWAS catalogue and OMIM – each serving human genetics in their own way*. dbSNP is a publicly available (at <http://www.ncbi.nlm.nih.gov/projects/SNP/>) collection of SNPs as the name suggests, but also includes indels and microsattelites [236-238]. As of January 2015 (current build version: 142), there are over 350 million submissions (known as ss#) just for the *Homo sapiens* build identifying over 100 million RefSNP clusters (known as rs#) with over 50 million of them validated (statistics from http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi). Being such a large resource of variants comes with its down sides, and dbSNP is no different (reviewed by Day, see reference [87]). Musumeci and colleagues estimated that over 8% of the SNPs deposited in dbSNP are amplification and/or sequencing artefacts which arise due to duplicated genes [239]. There are obvious caveats with using this database as a resource to filter out variants in order to distinguish neutral ones from pathogenic ones, as the latter could be in the database due to (human or machine) error. One way to go about solving this problem in Mendelian disease studies could be to use only the common variants (e.g. >1% at population level) which are less likely to be highly penetrant; and using either Ensembl VEP or the UCSC genome browser (<http://genome-euro.ucsc.edu/>) allows the filtering of these common variants with relative ease. ClinVar is another freely available mutation database which aims to collect all clinically relevant variants reported in the literature. HGMD (Human Genome Mutation Database) aims to collect *all* mutations reported the literature. The ‘Public’ version of HGMD can be incorporated freely via Ensembl VEP, however it is 3 years out of date.

Besides the abovementioned, there are many other forms of variant collections such as ‘locus specific databases’, which aim to collect and visualise sequence variant at the gene level. Since there is a drastic increase in genetic data density, there is also a concomitant demand for specialist databases as the sheer size of data that is available is becoming unmanageable for many researchers.

* The GWAS catalogue includes all SNPs which have been associated with a phenotype by a GWAS, whereas OMIM includes all variants which are associated with Mendelian disorders.

1.7.2. Clinical uses

Understanding the genetic basis of Mendelian and common complex disorders has allowed the development of numerous diagnostic tests, and initiation of many preventive interventions and curative therapies. Projects such as the HapMap project facilitated the foundation of several international consortia which have combined the data from multiple institutes and performed highly powered analyses of complex traits – especially in the form of GWAS. The studies have identified many novel loci associated with traits such as obesity, bone mineral density and cardiovascular diseases. As of Jan 2015, there are over fourteen thousand SNPs associated with hundreds of traits in the GWAS catalogue (<http://www.genome.gov/gwastudies/>). Many of the genes in the same LD block as the SNPs found to be associated with these traits are also known to harbour mutations in rare monogenic disorders [240]. Genetic information obtained at these loci can be used in clinical settings and at a public level for prognosis and prevention. Other immensely ambitious initiatives such as The International Rare Disease Research Consortium aim to develop tools for diagnosis of every single known rare disease and therapies for a couple of hundred of them [241].

1.7.3. Bioinformatics uses

The availability of publicly accessible genetic data from various populations has enabled the development of reliable bioinformatics software. Some of the available databases are used to ‘train’ and develop algorithms for mutation effect prediction software (e.g. FATHMM [242], SIFT [243], Polyphen-2 [244]) and to test the reliability of tools by enabling comparison of the different tools available. The predictions from bioinformatics tools are helpful in deciding which variants are likely to be causal and allow a systematic way of ranking them.

SNPs that have been identified to be highly correlated (i.e. in the same LD block) in projects such as the HapMap project, have enabled researchers to statistically infer

(or impute) additional loci. Large scale studies have only been made feasible with the advancements in bioinformatics tools as well as advancements in genetics. This is because gigantic amounts of data are produced from NGS technologies, and automated methods are needed for base calling, read alignments, variant calling (with respect to a reference sequence), annotation and filtering. The output cannot be analysed manually either, thus additional bioinformatics tools are required in the visualisation, filtering/ranking and interpretation stages too. Each step requires vast amount of theory and the implementation of these theories into code, and then to a user friendly package for others to use.

1.8. Summary of Aims and Objectives

This thesis had several aims; and the results of the analyses carried out with respect to achieving each aim was presented within a chapter of its own for clarity.

The main aim of Chapter 3 was to provide a generic guide on how to identify highly-penetrant disease causal mutations from NGS data obtained from consanguineous and outbred individuals. Chapter 3 therefore presents a (non-systematic) review of the relevant literature with regards to analysing NGS data, and the bioinformatics tools and databases that are available. Another key aim was to publish a review paper in a peer-reviewed journal using the results arising from this chapter.

Building on Chapter 3 by using the relevant tools described there, Chapters 4 and 5 aimed to identify Primary Ciliary Dyskinesia (PCD) and Autosomal Recessive Intellectual Disability (ARID) causal genes/variants/regions in consanguineous offspring respectively. Chapter 4 presents the results of analyses from 6 different families who have at least one PCD affected member; and similar to Chapter 4, Chapter 5 presents results from a (single) large consanguineous family which had 6 ARID affected members. Another key aim was to publish the novel findings arising from these analyses in peer-reviewed journals and academic conferences.

Chapter 6 aimed to further utilise the already available whole-exome sequencing (WES) data obtained from one of the consanguineous families in Chapter 4 by

carrying out a 'proxy molecular diagnosis' and identifying the disease causal variant in other members of the same family (affected from a different disease). Another key aim was to write a paper(s), reporting the findings of this chapter and publish it in a peer-reviewed journal.

Finally, Chapter 7 aimed to carry out a (non-systematic) review on the relevant literature with regards to the importance of consanguinity and consanguineous populations to human genetics, and to put forward arguments on the use of consanguineous populations 'as a whole' rather than only 'cherry-picking' certain families with disease. Similar to previous chapters, an important aim of this chapter was to publish a review paper in a peer-reviewed journal.

In addition to the above, each chapter has its own 'Aims and Objectives' section where the above is re-iterated and/or further detailed with respect to each chapter's aims.

CHAPTER 2. OVERVIEW OF METHODS

In this chapter I have presented the methods and tools (both wet-lab and computational) frequently used across the different chapters of the thesis such as DNA extraction, PCR, variant calling and mutation effect prediction. Other specific methods were detailed in their respective chapters. Where appropriate, the theory behind the methods was explained in detail. The ethics behind the studies carried out in this thesis was also included in this chapter.

2.1. Materials/samples

Participants

All nine subjects (from 6 different families) studied in this thesis were of Saudi Arabian ancestry. Study relevant details (e.g. phenotype, family members) about the subjects can be found in section 4.4.2. The recruitment process can be found in section 2.2.

Blood samples

50ml of peripheral blood was collected from the participants and stored in EDTA buffer (pH 8.0) at 4C° (and equilibrated to room temperature) before being used for DNA extraction. Care was taken to extract DNA from the samples as soon as they were received. Remaining samples were discarded after use.

Buccal swab samples

Buccal swab samples were collected using Oragene OG-575 (DNA Genotek Inc., Kanata, Ontario, KLK 1L1, Canada) saliva collection kits and kept at room temperature before DNA extraction. Remaining samples were kept for a year (maximum) before removal.

2.2. Ethics

For the family based studies, ethical approval was obtained from the King Saud University/King Khalid Hospital, Riyadh ethics committee (approval number: E-11-448). Family and individual consent was written, with the recognition that positive findings would be diagnostically reconfirmed in conjunction with clinical counselling and feedback. Consent was obtained following a family/patient information session (explained in advance via phone conversation, then recapped in clinic). Concerning minors, written parental consent was obtained from both parents, unless stated otherwise. The record of family visit to clinic was also added to hospital clinical notes as was the record of any buccal samplings for DNA agreed and undertaken.

For the mutation screening sample of 256 individuals, participation was voluntary and informed written consent for anonymised genetic studies was taken in keeping with King Saud University College of Applied Medical Sciences guidelines. The local population reference DNA sample comprised of male and female student volunteers of Saudi Arabian ancestry studying at the King Saud University (Riyadh, Kingdom of Saudi Arabia).

The University of Bristol and the Bristol Genetic Epidemiology Labs – where these studies were undertaken – holds Human Tissue Authority (HTA) licences and all relevant samples were declared to the HTA before receipt of the samples (as requested by law). Additionally, I received training on Research and human tissue legislation from an online course designed by the MRC (details at: <http://www.rsclearn.mrc.ac.uk/>). The certificate confirming training can be found in section 9.1 (Figure 10.2).

No data was uploaded to external online servers nor shared with any other research group to ensure confidentiality of the participants' data. Also, to increase anonymity of the participants, some of the figures relating to genome-wide data have been edited without affecting the regions/parts related to the results presented.

The ethics statements here encompass all studies carried out throughout this thesis unless stated otherwise.

2.3. Wet-Laboratory methods

2.3.1. DNA extraction and quantification

DNA from Blood

DNA was extracted from peripheral blood (stored at 4C°) using the QIAamp DNA Mini kit (Catalogue No: 51304) provided by QIAGEN Ltd (Manchester, M15 6SH, United Kingdom); and the protocol for DNA Purification from Blood or Body Fluids was followed in the QIAamp DNA Mini and Blood Mini handbook (available online at: <http://www.qiagen.com/products/catalog/sample-technologies/dna-sample-technologies/genomic-dna/qiaamp-dna-mini-kit#resources>). 400µl of blood was used in the extraction process, thus numbers in the protocol were adapted accordingly (i.e. doubled). All blood samples were equilibrated to room temperature (~20 C°) before extraction.

DNA from Buccal swabs

DNA was extracted from samples using DNA Genotek prepIT-L2P (Catalogue No: PT-L2P); and the prepIT laboratory protocol for manual purification of DNA from 0.5mL of sample was followed.

DNA storage

All extracted DNA were dissolved in TE buffer (10mM Tris-HCl, 1mM EDTA, pH 8.0) and stored at -20C°. Aliquots were kept to ensure less degradation due to thawing and re-freezing.

DNA quality and concentration quantification

DNA concentration was measured using the Invitrogen (Fisher Scientific - UK Ltd, Loughborough, LE11 5RG, United Kingdom) Qubit® dsDNA HS Assay Kit (Catalogue No: Q32851) following the manual MAN0002325 (URL: <http://tools.lifetechnologies.com/content/sfs/manuals/mp32850.pdf>).

Discrepancies between NanoDrop and Qubit Fluorometer

Discrepancies were noted between the NanoDrop and Qubit fluorometer methods available when DNA concentrations of the same samples were measured (Tables 2.1 and 2.2).

Sample Name	Concentration (ng/μL)	Volume (μL)	Total quantity (μg)
1	130.7	21	2.75
2	279.9	9.8	2.75
3	85.9	32	2.75
4	105.2	26.1	2.75
5	165.6	16.6	2.75
6	260.3	10.6	2.75
7	61	45.1	2.75
8	76.8	35.8	2.75
9	343.5	8	2.75

Table 2.1 NanoDrop results for randomly selected 9 samples

Sample Name	Concentration (ng/μL)	Volume (μL)	Total quantity (μg)
1	65.4	22	1.44
2	70.6	24	1.69
3	50.5	25	1.26
4	51.7	28	1.45
5	17.4	27	0.47
6	80.8	24	1.94
7	50.5	38	1.92
8	53.3	29	1.55
9	41.3	24	0.99

Table 2.2 Qubit Fluorometer results for the same 9 samples (i.e. Sample 1 in Table 2.1 refers Sample 1 in Table 2.2)

DNA concentration measurements were higher for all samples when the NanoDrop kit was used. The difference was as high as 5 times; and 2 times for most in relation to the Qubit fluorometer kit used.

As BGI-Tech – where the whole-exome sequencing in this thesis was carried out - were using and also recommending Qubit fluorometer, this technique was used throughout this thesis hereafter to avoid these discrepancies between the two techniques.

Testing DNA sample quality

The below grades were stated in the methods section of the chapters where WES data was used (e.g. Chapter 4). Where DNA quantity was deemed 'sufficient' but the DNA concentration was not, the SpeedVac DNA concentrator (SPD2010, Fisher Scientific - UK Ltd, Loughborough, LE11 5RG, United Kingdom) machine was used to increase the latter.

DNA quality grading by BGI-Tech was as follows (adapted):

Level A meant that the sample was of highest quality, and the amount of sample was sufficient for two or more library constructions. Level B meant that the sample was of high quality, but the amount of sample only satisfied one library construction. Level C meant that the sample did not fully meet the requirements of library construction and sequencing. Level D meant the sample did not meet the requirements of library construction and sequencing. Level C samples had the risk of failure in library construction and may have failed in sequencing because of the low yield of library. Data quality of sequencing could also have been affected. For example, there may have been preferential amplification, high duplication rates, low coverage and/or abnormal GC contents. Where a sample was deemed to be in the Level C category, the amount of sample sent was increased. Level D samples could not be quantified accurately and therefore had the risk of failure in library construction and sequencing, because of low yield of library. Data quality of sequencing could also have been affected. The other abovementioned problems could also occur with these samples. Where a sample was deemed to be in the Level D category, a new sample was resent.

These grades were used to rate the quality of the DNA samples used for WES (see section 4.5.1).

2.3.2. Polymerase Chain Reaction (PCR)

Polymerase chain reaction (widely known by its abbreviation, PCR) is a technique used to amplify a region of interest from a single copy of DNA [207]. Since its discovery in 1985, PCR has been used as the underlying method for many different purposes such as mutation screening, transcript quantification and forensic purposes (e.g. real-time PCR, ARMS-PCR).

The PCR technique capitalises on the amplification and heat withstanding abilities of the DNA polymerase enzyme (known as Taq polymerase) synthesized by *Thermus aquaticus*. Two primers (named forward and reverse) specifically designed for the region of interest is allowed to anneal (or hybridise) to a previously melted (unzipped) DNA strand; and by intricate changing of the environment temperature, the polymerase enzyme is manipulated to start adding dNTPs to extend the 3' end of the primers (Figure 2.1). This in turn ultimately amplifies the region between the two designed primers.

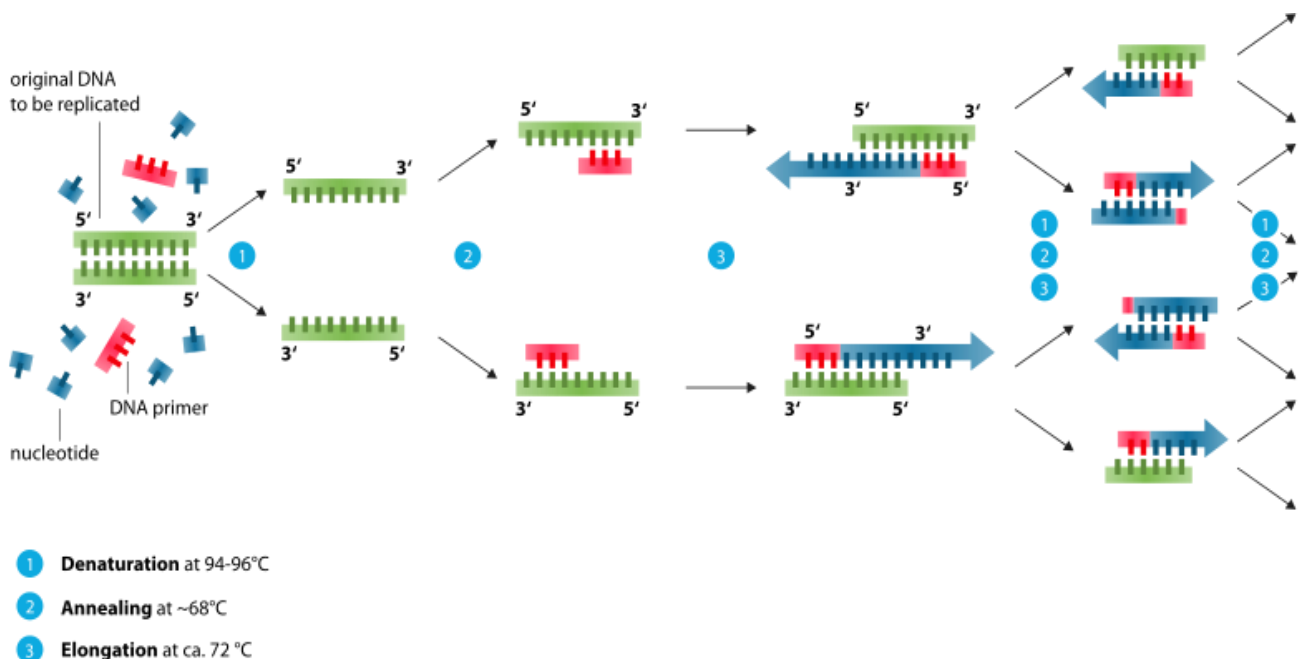


Figure 2.1 Schematic representation of the PCR cycle. The numbers in blue circles represent stages of PCR. The number of copies is doubled with each PCR cycle. Image reproduced under the creative commons licence, URL: http://en.wikipedia.org/wiki/Polymerase_chain_reaction).

Primer designing for PCR

The following represents a generic version of the method used throughout this thesis for designing a primer for PCR (unless stated otherwise). This method was used to amplify a region of interest (see respective sections for the regions); and the resulting amplicons were then sent to and sequenced by a sequencing company such as GATC-Biotech (Köln, Germany). GATC-Biotech offered Sanger sequencing (called SUPREMERUN) for PCR amplicons and this service was used throughout unless stated otherwise.

(i) To start designing a primer, I clicked on the link below:

Primer Blast (link: <http://www.ncbi.nlm.nih.gov/tools/primer-blast/>)

(ii) In the 'PCR Template' section at the top, I entered the 'Accession ID' of my transcript of interest from RefSeq (link: <http://www.ncbi.nlm.nih.gov/refseq/>) if working with mRNA; but if I was interested in amplifying a genomic region, then I used Ensembl (link: <http://www.ensembl.org/index.html>, searched for my gene of interest in the Ensembl homepage, clicked on gene of interest in the results, then in the 'Gene' view, clicked on the 'Sequence' in the 'Gene-based display' on the left, and then copied the 'Marked-up sequence' in FASTA format and pasted it into the PCR Template).

I then calculated where my variant of interest was located in the FASTA sequence (Ensembl) or in the transcript (RefSeq mRNA) that I copy-pasted and filled in 'Forward Primer' and 'Reverse Primer' accordingly (leaving 150bp around my variant on both sides – e.g. if my variant was located at position 500 in my FASTA sequence, then I typed 350 into 'From' in 'Forward Primer' and 650 into 'To' in 'Reverse Primer', leaving the other two boxes empty).

(iii) In 'Primer Parameters': To get the amplicon sent and sequenced at a company, I kept the PCR product size manageable for sequencing (e.g. 150bp to 300bp).

(iv) When working with human genomic data, I changed 'Database' to 'Genome (reference assembly from selected organisms)' and selected 'Homo sapiens' as 'Organism' in 'Primer Pair Specificity Checking Parameters'.

I then clicked Advanced parameters.

(v) I then changed 'Primer Size' in 'Primer Parameters' to 18 (min), 22 and 25 (max) respectively.

Then changed 'Primer GC Content (%)' to 40.0 and 60.0 respectively.

Then changed 'GC Clamp' to 1.

Then changed 'Max Poly-X' to 3.

Then ticked the 'SNP handling' box.

(vi) I scrolled to the bottom and clicked 'Show results in a new window' before clicking 'Get Primers'.

(vii) After the results were outputted, I selected a **couple*** (designing two in case one fails) of primer pairs and tested them in an *in-silico* PCR software (e.g. UCSC In-Silico PCR, link: <http://genome.ucsc.edu/cgi-bin/hgPcr>).

If the amplicon produced in the *in-silico* PCR program contained my loci of interest (checking that my variant[†] of interest is located towards the centre of the amplicon), I then checked for hairpin formation (both for forward and reverse primers) using software such as OligoCalc (link: <http://www.basic.northwestern.edu/biotools/oligocalc.html>).

* I also checked that the GC content of the forward and reverse primers were similar to each other for each primer pair.

[†] If my gene of interest was on the reverse strand, then a software such as Reverse Complement was used to change the sequence of my amplicon to its complement sequence so that it matches the Ensembl gene sequence that I was looking for (obtained the sequence in FASTA format in step ii)

Once the primer pairs passed all these tests, I have ordered them from a company called Eurofins (Ebersberg, Germany).

When choosing annealing temperature (T_a) for my primers, (for primers with no unintended* targets) I usually set them 6-7 degrees Celsius below the melting temperature (T_m) of the primer with lowest T_m . However T_a and $MgCl_2$ gradients (i.e. titration) were also tried if PCR did not work for both the primers designed.

If conventional Taq polymerase did not work (or produces too many unwanted targets), I tried using Hot Start activated polymerase.

Using PCR-based methods for variant screening

Although PCR was mainly used for amplifying certain regions of the genome, combining PCR with specifically designed primers proved to be a cost-effective way to genotype SNV/SNPs. Throughout this thesis, two different methods were successfully used to genotype SNVs: ARMS-PCR (Chapter 6) and PCR-RFLP (Chapter 4). Details on the methods are as follows:

ARMS-PCR

ARMS-PCR is a PCR based technique used to detect known point mutations and perform mutation screening in large samples (a variant of the original method by Newton *et al* [245]). ARMS-PCR makes use of the fact that primers hybridise with complementary regions in the genome thus by introducing an allele specific base at the 3' end of a primer, the primer will hybridise only when the allele is present. However since primers may hybridise to none-perfect-matching regions in certain environments, a secondary mismatch is also presented to the primer enabling a more specific test for SNVs. The theory behind ARMS-PCR is depicted in Figure 2.2 below.

* If there are other bands in the gel, I tried increasing T_m as this allowed the primer to hybridize to the perfectly matching DNA sequence and not to the other unintended regions

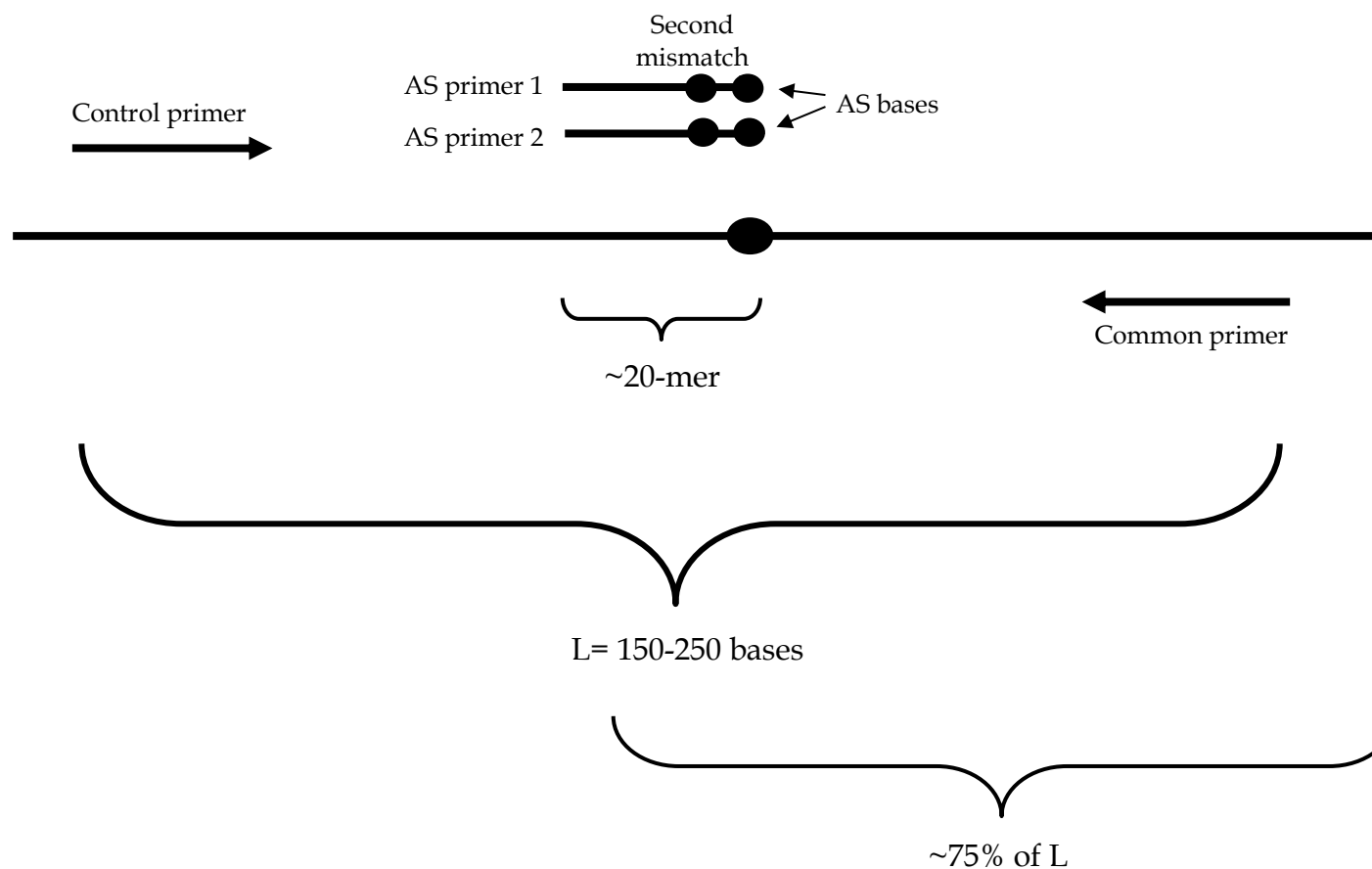


Figure 2.2 Schematic of the ARMS-PCR methodology. Four primers are designed (incl. 2 allele-specific primers) and PCR is carried out using the control and common primers together with one of the AS primers. PCR is then repeated with the other AS primer. The resulting amplicons are then electrophoresed to observe the presence/absence of the mutant allele in the subjects analysed.

PCR-RFLP

PCR-RFLP is another PCR based technique used to detect targeted point mutations. PCR-RFLP combines PCR with the conventional restriction fragment length polymorphism (RFLP) technique where a restriction enzyme is used to digest fragments (i.e. amplicons) containing a certain palindrome which the chosen enzyme recognises. Thus, with a custom-picked enzyme (where feasible), PCR fragments with the mutation of interest can be digested allowing screening for the variant in a systematic and automatic manner.

2.3.3. Gel electrophoresis and visualising PCR products

Once PCR was finished, it was crucial to check whether the process had worked successfully. Therefore a small proportion of the PCR product (usually 10µl throughout this thesis, unless stated otherwise) was mixed with a dye (which illuminates when UV light is shone at it) and pipetted in to a well. These products were then electrophoresed on an agarose gel for 90 minutes (unless stated otherwise) which was then viewed on a camera (placed on top of an UV trans-illuminator) via a PC software (Remote Capture DC was used, link: <http://canon-utilities-remotecapture-dc.updatestar.com/>).

Throughout this thesis, I have mixed 10µl of PCR product and 5µl the Promega (Promega UK Ltd, Southampton, SO16 7NS, United Kingdom) 100bp Ladder (Catalogue No: G210A) with 2µl and 1µl of Promega Blue/Orange loading dye 6X (Catalogue No: G1881) respectively. The agarose gel concentration was 1.5% and the number of wells made was either 10 or 20 (visible on photos presented). Unless stated otherwise, these were the solutions and amounts used.

96-well MADGE

Microplate Array Diagonal Gel Electrophoresis (MADGE) is a spin on the traditional 10-well agarose gel electrophoresis system. MADGE is based on a standard 96-well microplate array format (9mm well to well distance and 2mm cubic wells where the array is rotated 71.6° to extend the track length to 26mm).

MADGE uses a plastic well former, into which an acrylamide gel was poured. A silanised glass plate was then placed on top. After the gel had set, the glass plate and attached gel were removed from the well former.

The whole process from making the gel to removal of glass plate took less than 30 minutes. Then, electrophoresis at 150V was carried out for 20 minutes.

Details on the ingredients and the protocol are available on the original paper by Day *et al* [246].

2.3.4. Exome targeting

As whole-exome sequencing was carried out at the BGI-Tech sequencing centre (Building No.11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China), the following exome targeting procedure was adapted from their 'Pipeline of Experiment' section:

Initially, qualified genomic DNA sample was randomly sheared into fragments with lengths between 150 to 200bp, and adapters were ligated to both ends of the resulting fragments. The adapter-ligated templates were purified using AgencourtAMPure SPRI beads and fragments with insert size of about 200bp were excised. Extracted DNA was amplified using ligation-mediated polymerase chain reaction (LM-PCR), purified and hybridized to the SureSelect Biotinylated RNA Library (BAITS) for enrichment. Hybridized fragments were then bound to streptavidin beads, and fragments that did not hybridise were washed out (after 24 hours). Captured LM-PCR products were checked for quality control using Agilent 2100 Bioanalyzer to estimate the magnitude of enrichment. Each captured library was then loaded onto the Illumina Hiseq2000 platform, and high-throughput sequencing was carried out for each captured library independently, to ensure that each sample met the desired average fold-coverage (50x). Detailed information about the Agilent SureSelect exome capture kit can be found at:

<http://www.halogenomics.com/sureselect/how-it-works>.

2.4. Bioinformatics and Statistical methods

2.4.1. DNA Sequencing and Mapping

Whole-exome sequencing

Throughout this thesis where WES was carried out, the exomes of the participants were captured using the Agilent SureSelect Human All Exon 50M exon capture kit (Agilent Technologies, Inc. Santa Clara, CA, 95051, USA) and WES data was obtained by subsequent sequencing using the Illumina Hiseq2000 platform (Illumina, Inc. San Diego, CA, 92122, USA). WES was carried out at the BGI-Tech sequencing centre.

Raw image files were processed by Illumina base calling Software 1.7 for base calling with default parameters and the sequences of each individual were generated as 90bp paired-end reads. The Burrows-Wheeler Aligner (BWA) [247] software was used to align the reads to the human genome reference sequence (i.e. hg19), filtering out reads which have extensive low base quality (more than half of the bases which have a base quality of ≤ 5 , including no calls) and/or with a mapping score of zero. Picard (<http://picard.sourceforge.net>) was used to sort and mark duplicated reads and the alignment results were generated in BAM format.

Sanger sequencing

Throughout this thesis where PCR amplicons required sequencing, Sanger sequencing was carried out at GATC-Biotech centre located in Köln, Germany. Data was received in AB1 format. The DNA chromatogram results were viewed using the Chromas Lite (v2.1.1, link: http://technelysium.com.au/?page_id=13) software.

2.4.2. Variant calling and annotation

Single nucleotide polymorphisms (SNPs) in the WES data were called using SOAPsnp [248] and small insertion/deletion events (indel) were detected by SAMtools (mpileup function). The variants were validated using GATK, and

exported in VCF format [249-251]*. VCF variant annotations were obtained from the Ensembl Variant Effect Predictor (VEP) [252] and ANNOVAR [253]. The UNIX commands and parameters used can be found in section 9.8.

CNVs were detected from the WES data using the Control-FREEC software (using Family 4 individual 1 as the control) [254].

Using Control-FREEC

To use Control-FREEC, the following files were needed: (i) FREEC_LINUX32.tar (ii) config file (see section 9.5.3 for one made for family 1 individual 1) (iii) hg19.len (iv) out100m2_hg19.gem (v) BAM file for sample/ query and control sequence (vi) 'exome captured regions' file e.g. truseq_exome_targeted_regions.hg19.bed.chr (vii) makeGraph.R

Files (i - iv) and (vii) were available to download from the Control-FREEC website (<http://bioinfo-out.curie.fr/projects/freec/>); and file (vi) was available at the Illumina website (URL: <http://support.illumina.com/sequencing/downloads.html>).

Executed command:

```
./freec -conf <config file>
```

Command used for creating genome-wide CNV graphs:

```
awk '$3!=-1 {print}' file.rmdup.bam_ratio.txt > file.rmdup.bam_ratio_noNA.txt
cat makeGraph.R | R --slave --args 2 file.rmdup.bam_ratio_noNA.txt
```

2.4.3. Mutation effect predictors

Throughout this work, I have used bioinformatics software to systematically prioritise certain variants over others depending on where they occur (e.g. candidate

* Genotype with the highest probability at a given locus was identified for each individual sample and a consensus sequence was assembled. Using the consensus sequence, variations between the identified genotype and the reference was saved as a VCF file.

gene, novel gene, intron), what their consequence was (e.g. indel, synonymous mutation, nonsynonymous mutation, frameshift) and their functional effects (e.g. deleterious, benign). The latter factor was still unknown for most variants thus mutation effect predictors which can achieve this in an automated fashion were needed to prioritise variants (according to their predicted functional effects). There are many types of these bioinformatics tools and each came with their own algorithms, methods and theory behind them. In this work, I have mainly used FATHMM, SIFT, PolyPhen-2 and Condel for predicting the effects of nsSNVs. The reasons behind these choices were their credibility (due to high citation rates), performance rates and relative ease of use (for a comprehensive review on the performances of these tools, see [255]). FATHMM was produced at the University of Bristol by Dr. Hashem A. Shihab (who sat next to me in the BGEL office) thus the availability of instant help was also a reason behind the use of this particular tool besides its high prediction performance. Also these tools were continually updated which is a must in the rapidly developing fields of bioinformatics and genetics. Details on each of these tools are below:

FATHMM

The Functional Analysis through Hidden Markov Models software (named FATHMM, read 'fathom' [242]) predicts the functional effects of nsSNVs by combining amino acid sequence conservation (using HMMs) with 'pathogenicity weights' which are derived from the relative frequencies of disease associated and functionally neutral amino acid substitutions mapping onto conserved protein domains. The tool is available for download as a standalone package at <http://fathmm.biocompute.org.uk/> but is also built in to the Ensembl VEP online tool as a 'plugin' (see <http://www.ensembl.org/info/docs/tools/vep/script/index.html> or section 9.8 in the Appendices chapter on how to use). I also wrote a python script which reformats missense mutations 'grep'ed out of a VEP file (using command: `grep missense file.vep`) to FATHMM input format (see `missense2fathmm.py` in section 9.8).

SIFT

Sorting Intolerant From Tolerant software (named SIFT [243]) is a highly used algorithm which makes use of position specific iteration BLAST (PSI-BLAST, developed by [256]) to calculate the probability of an nsSNV to have a deleterious (functional) effect on protein function. The deleteriousness score is conditioned on the most frequent amino acid residue having no negative impact on protein function, thus the algorithm is essentially making use of information about how conserved an amino acid is. The tool is available to download from <http://sift.jcvi.org/> but is also built in (using --sift command) to the Ensembl VEP online tool (see <http://www.ensembl.org/info/docs/tools/vep/script/index.html> or section 9.8 in the Appendices chapter).

PolyPhen-2

Polymorphism Phenotyping v2 (named Polyphen-2 [244]) is a trained algorithm which makes predictions on the impact that an amino acid substitution may have using structural (e.g. DNA sequence and protein structure) as well as evolutionary comparative (e.g. conservation) data. The tool is available to download from <http://genetics.bwh.harvard.edu/pph2/> but is also built in (using --polyphen command) to the Ensembl VEP online tool (see <http://www.ensembl.org/info/docs/tools/vep/script/index.html> or section 9.8 in the Appendices chapter)

Condel

Consensus Deleteriousness of nsSNVs tool (named Condel [257]) combines the outputs of five tools (including SIFT and PolyPhen-2) and computes a weighted average of the scores; and in theory Condel outperforms the other prediction tools when they are used independently [257]. However, one down side of Condel is that it does not provide predictions for variants where the other tools do not provide a prediction. The tool is available for download at <http://bg.upf.edu/condel/home>

but is also available as a ‘plugin’ to the Ensembl VEP online tool (see <http://www.ensembl.org/info/docs/tools/vep/script/index.html> or section 9.8 in the Appendices chapter).

Condel has now been succeeded with Condel-2 which combines FATHMM with MutationAssesor to make the predictions.

2.4.4. Candidate genes/variants

As WES enables a ‘hypothesis free’ search for a causal variant, even genes which have no prior connection with disease of interest analysed are sequenced; and variants within these genes are detected. Thus, usually a list of candidate genes is created to promote (rank higher) variants which fall into this list.

Throughout this thesis, I have used the literature as a way of gathering a list of candidate genes. Animal knockout studies and proteomic analyses were paid special attention. Besides the literature, software predicting protein-protein interactions such as STRING [258], and disease/pathway specific databases (e.g. Ciliome database – see Table 10.2 for Ensembl IDs of all genes [259]) were also used.

For variants, HGMD and ClinVar were checked to see whether they were reported previously. Additionally for variants which have not been reported, homology BLAST was used to view how conserved a certain residue was, which may indicate functionality.

2.4.5. Protein structure modelling

Although it is very hard to reliably predict the structure of a protein from its amino acid sequence at current, where deemed applicable, the normal and mutant versions of the proteins were modelled using the Robetta software [260].

Robetta (<http://robetta.bakerlab.org/>) is a full-chain protein structure prediction server, parsing protein chains into putative domains using the Ginzu* protocol

* Ginzu is a Protein Databank (PDB) template identification and domain prediction protocol that attempts to determine the regions of a protein chain that are aligned to PDB templates with

(modelling domains using homology or *ab initio* modelling). The resulting files were viewed using the RasWin software (www.rasmol.org).

2.4.6. Calculating F values and identifying autozygous regions

Calculating the inbreeding coefficient (i.e. F) of an individual can provide essential information, especially in studies of consanguineous individuals and/or populations. The F value will indicate the amount of autozygosity that is expected in an individual, which will then have implications on where the (autosomal recessive disease) causal variants may lie. Sewall Wright published a landmark paper in 1922 where he produced a simple formula for calculating the inbreeding coefficient of an offspring (F_o) where the parents are related (Equations 2.1 and 2.2) [261].

Wright's formula is still used today for less complex pedigrees. I have also used this formula where necessary. For more complex analyses or when familial data is not very reliable, long runs of homozygous stretches of DNA (LROH) were analysed in the offspring to quantify autozygosity (i.e. F value); and software such as AutoZplotter (which takes VCF files as input), David Pike's method (designed specially for 23andme SNP chip array data) and/or Plink were used.

$$F_o = \sum \left(\frac{1}{2}\right)^{n+n'+1}(1 + F_a)$$

Equation 2.1 Calculating the inbreeding coefficient of an individual (F_o). n and n' represent the number of generations from father and mother respectively to their common ancestor. If the common ancestor is also an inbred himself/herself, then the inbreeding coefficient for the ancestor (F_a) must also be worked out from his/her own pedigree.

reasonable confidence. In regions where templates are not detected, it attempts to identify domains – see Table 10.3 for details on PDB file format. The most likely structure is chosen by comparison with the templates matching the query sequence and the energy levels of the final protein structure.

$$F_O = \sum \left(\frac{1}{2}\right)^{n+n'+1}$$

Equation 2.2 Simpler version of Wright's inbreeding coefficient formula. This version can be used when the common ancestor is not an inbred himself/herself, making the calculations much simpler as $F_a = 0$, therefore making $(1 + F_a) = 1$ which makes this part of the formula redundant.

Details on how to quantify autozygosity and which software are available can be found in section 3.3.3 (under 'Autozygosity mapping'). The expected and observed F values can be different from each other due to recombination which occurs during meiosis. Therefore where applicable, both expected (calculated from family pedigree) and observed (calculated by identifying total LRoH within an individual's genome and dividing by total genome size i.e. 3.3×10^9 bp) F values were presented. If the sex chromosomes were not included, then the total genome size will be set as 3.1×10^9 bp (subtracting the total of the lengths of the two sex chromosomes, setting X's length as 150Mb and Y's length as 50Mb).

2.5. Literature reviews

In addition to the introduction chapter (i.e. Chapter 1), literature reviews were carried out in Chapters 3 and 7. The literature reviews were mostly based on key references and reviews in the respective areas (not selected systematically, but through experience of myself and supervisors in the respective fields); then, where necessary, by branching out to the references cited in these papers. As the bioinformatics field (e.g. new tools, new databases) is changing all the time, new tools were paid special attention to provide updates on tools/datasets that were already being used. Where (a section of) the chapter was published, the reviewers' suggestions have also helped shape the respective chapter/section.

As for the literature reviews carried out in Chapters 4 and 5, all papers matching the keywords "consanguineous", "autosomal recessive" and the respective disease name (e.g. PCD, Intellectual Disability) has been screened for relevant information.

CHAPTER 3. IDENTIFYING HIGHLY-PENETRANT DISEASE- CAUSAL MUTATIONS FROM NEXT-GENERATION SEQUENCING DATA

Recent technological advances have created new challenges for geneticists and a need to adapt to a wide range of bioinformatics tools and an expanding wealth of publicly available data (e.g. mutation databases, software). This wide range of methods and a diversity of file formats used in sequence analysis is a significant issue, with a considerable amount of time spent before anyone can even attempt to analyse the genetic basis of human disorders. Another point to consider is that although many possess “just enough” knowledge to analyse their data, they do not make full use of the tools and databases that are available and also do not fully understand how their data was created. The primary aim of this chapter was to document some of the key approaches and provide an analysis schema to make the next-generation sequencing (NGS) data (see section 1.6.2 for details on NGS) analysis process more efficient and reliable in the context of discovering highly-penetrant disease-causal mutations. In this chapter, I have also compared the methods used to identify highly penetrant variants when data is obtained from consanguineous individuals as opposed to non-consanguineous; and when Mendelian disorders are analysed as opposed to common-complex disorders.

This chapter was a stand-alone review which was an aim and an output of this thesis (see section 1.8). This chapter was a lynchpin for Chapters 4, 5, 6 and 7 as the methods used here apply to the analyses conducted (or proposed to be conducted) in

these chapters. Also, as WES data was used throughout this thesis, the figures (and examples) provided here has reflected this.

3.1. Introduction

Next generation sequencing (NGS) and other high throughput technologies have brought new challenges concomitantly. The colossal amount of information that is produced has led researchers to look for ways of reducing the time and effort it takes to analyse the resulting data whilst also keeping up with the storage needs of the resulting files – which are in the magnitude of gigabytes each. The recently emerged variant call format (VCF) has somewhat provided a way out of this complex issue [251]. Using a reference sequence and comparing it with the query sequence, only the differences (i.e. variants) between the two are encoded into a VCF file. Not only are VCF files substantially smaller in size (e.g. for whole-exome data, <300x in relation to BAM files which store all raw read alignments), they also make the data relatively easy to analyse since there are many bioinformatics tools (e.g. annotation, mutation effect prediction) which accept the VCF format as standard input. The Genome Analysis Toolkit (GATK) made available by the Broad Institute also provides useful suggestions to bring a universal standard for the annotation and filtering of variants in VCF files [249]. The abovementioned reasons have made the VCF the established format for the sharing of genetic variation produced from large sequencing projects (e.g. 1000 Genomes Project, NHLBI Exome Project - also known as EVS). However the VCF does have some disadvantages. The files can be information dense, initially difficult to understand and parse. Comprehensive information about the VCF and its companion software VCFtools [251] are available online (vcftools.sourceforge.net).

Because of the substantial decrease in the price of DNA sequencing and SNP chip arrays [262], there has been a sharp increase in the number of genetic association studies being carried out, especially in the form of genome-wide association studies (GWAS, statistics available at www.genome.gov/gwastudies/). As whole genome sequencing (WGS) is prohibitively expensive for large genetic association studies

[263-265], whole exome sequencing (WES) has emerged as the attractive alternative – where only the protein coding region of the genome (i.e. exome) is targeted and sequenced [266]. The decision to carry out WES over WGS is not solely influenced by the cost which currently stands at one-third in comparison [267], but also by the fact that most of the known Mendelian disorders (~85%) are caused by mutations in the exome [216]; and reliably interpreting variation outside of the exome is still challenging as there is little consensus on interpreting their functional effects (even with ENCODE data [268] and non-coding variant effect prediction tools such as CADD [269], FATHMM-MKL [270] and GWAFA [271]). For complex diseases, WES can provide more evidence for causality compared to GWAS - assuming that the causal variants are exonic. This is because the latter uses linkage disequilibrium (LD) patterns between common markers [186] whereas WES directly associates the variant itself with the trait/disorder. Therefore using GWAS, especially in gene-dense regions, one cannot usually make conclusive judgements about which gene/variant(s) is causal without further sequencing or functional analysis.

WES has been successfully used in identifying and/or verifying over 300 causal variants for Mendelian disorders (statistics from omim.org/) (also see references [59, 272] for comprehensive discussion of the use and benefits of WES in clinical genetics). WES currently stands at approx. \$1000 for 50x read depth (variable prices, less for larger studies). However since there is a great deal of variation in the human genome [208], finding the causal variant(s), especially ones with low penetrance, is not going to be trivial. This problem can be exacerbated by the nature of the disorder(s) analysed. It is relatively easier to map variants causing rare monogenic diseases (when several affected individuals/families are available for analysis), as there is most likely to be a single variant present in the cases that is not in the controls; but in contrast, common complex (polygenic) disorders are much harder to dissect when searching for causal variants.

3.2. Aims and Objectives

In this chapter, my aims were to provide a guide for genetic association studies dealing with sequencing data to identify highly penetrant variants and compare the different approaches taken when data is obtained from unrelated and consanguineous individuals. Another aim was to make suggestions about how to rank single nucleotide variation (SNV) and/or insertion/deletions (indels) following the standard filtering/ranking steps if there are several candidate variants – using annotated variants within VCF files as examples. To aid the process of analysing sequencing data obtained from consanguineous individuals, I have also made available an autozygosity mapping algorithm (named AutoZplotter, script can be found in section 9.8) which takes VCF files as input and enables manual identification of regions that have longer stretches of homozygosity than would be expected by chance.

3.3. Methods

3.3.1. Stage 1 - Quality control & Variant calling

Before any genetic analysis, it is important to understand how the raw data were produced and processed to make better judgements about the reliability of the data received. Thorough quality control steps are required to ensure the reliability of the dataset. Lack of adequate prior quality control will inevitably lead to loss of statistical power; and increase false positive and false negative findings. Fully comprehending each step during the creation of the dataset will have implications on the interpretation stage, where genotyping errors (also known as ‘phantom’ mutations [273]) may turn out to be statistically associated (e.g. batch effects between case and control batches) or the causal variant may not be identified due to poorly applied quality control (QC) and/or filtering methods. The most fitting example for this comes from a recent Primary ciliary dyskinesia (PCD) study [217], where the causal variant was only detected after the authors manually noticed an absence of

reads in the relevant region of the genome (personal communication with authors). The subsequent variant was not only missing in the VCF files, but also in the initial BAM files - requiring remapping of reads. Another point of consideration from this finding would be that the authors knew where to look because the *RSPH9* gene (and the p.Lys268del mutation) was one of their *a priori* candidates [274]. This is also an example demonstrating the importance of deep prior knowledge and screening for known variants as it is impossible for one to manually check the whole exome (or the genome) for sequencing and/or mapping errors.

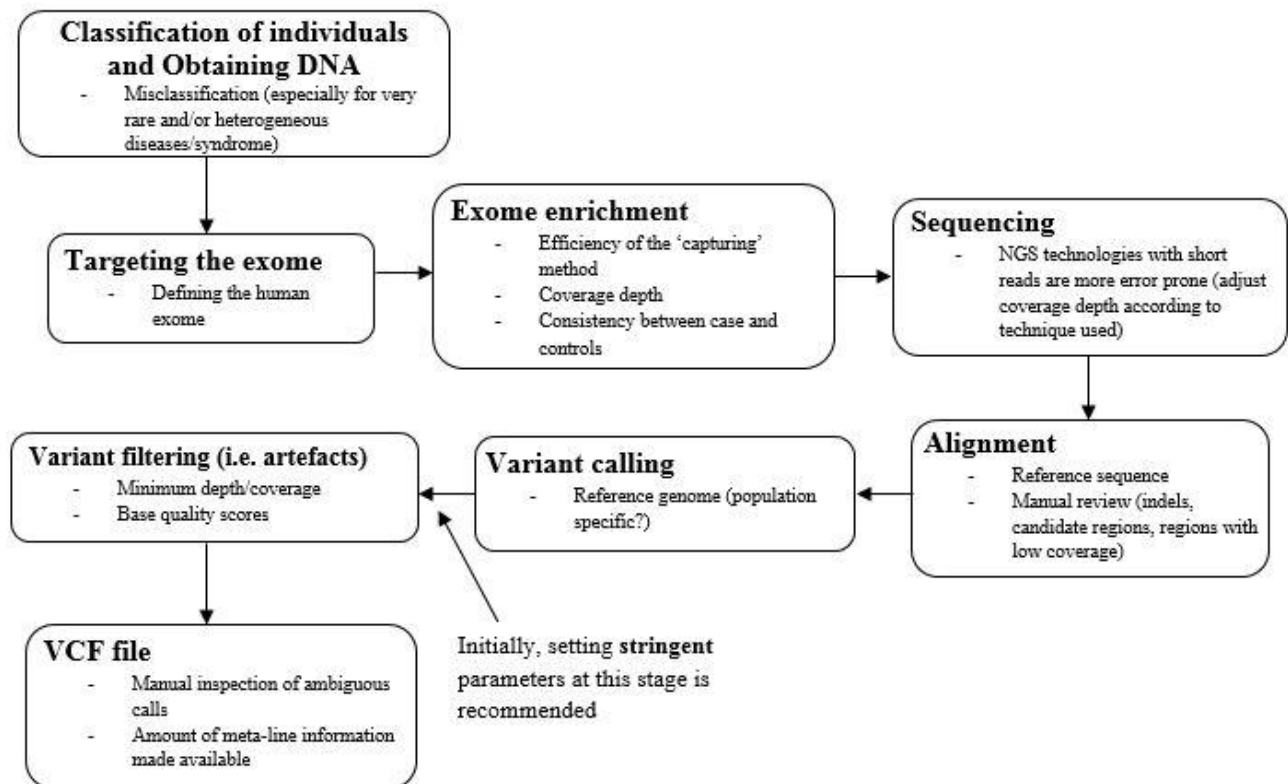


Figure 3.1 Steps in whole-exome sequencing. Understanding how the VCF file was created is important, as it can give an idea about where something may have gone wrong. The stages proceed from top to bottom and we have proposed ‘consideration points’ for each step (below each title).

Targeted sequencing

As far as WES projects are concerned, questions about coverage arise right from the start (Figure 3.1). Since knowledge concerning exons in our own genome is far from complete, there are differing definitions about the human exome coordinates.

Therefore, the targeted regions by the commonly used and commercially available Agilent SureSelect [275] and the Nimblegen SeqCap EZ [276] exome capture kits are

not entirely overlapping [277]. Thus it is possible that the missing regions of the exome due to the chosen probe kit may turn out to have the functional region in relation to the disorder analysed. One must also bear in mind that the kits available for targeting the exome are not fully efficient due to a certain quantity of poorly synthesized and/or designed probes not being able to hybridize to the target DNA. Next step is target enrichment where high coverage is vital as NGS machines produce more erroneous base calls compared to other techniques [278]; therefore, especially for rare variant analyses, it is important to have data with high average read depth (e.g. $\geq 50\times$).

Mapping sequence reads

The raw reads produced should then be aligned to a reference genome (e.g. GRCh38 – see NCBI Genome Reference Consortium) and there are many open source and widely applied tools available for this purpose (Table 3.1). However, solely depending on automated methods and software can leave many reads spanning insertions and deletions (indels) misaligned, therefore post-reviewing the data for mismapping is always a good practice, especially in the candidate regions. Attempting to remap misaligned reads with a lower stringency using software such as Pindel would be an ideal way to go about solving such a problem [279]. GATK also provides a base recalibration and indel realignment algorithm for this purpose.

Effective variant calling depends on accurate mapping to a dependable reference sequence. If available, using a population specific reference genome would be most ideal to filter out known neutral SNPs existing within the region of origin of the analysed subjects (e.g. East-Asian reference genome for subjects of Japanese origin). Inclusion of ambiguity codes (e.g. IUPAC codes) for known poly-allelic variants to create a composite reference genome can also be useful (although not essential).

Name	References	Notes
BFAST	[280]	These aligners use similar algorithms to determine contiguous sequences however MAQ and BWA are widely used and have been praised for their computational efficiency and multi-platform compatibility [82]. Bowtie is available with Galaxy servers which aims to provide push-button bioinformatics for users who are not familiar with UNIX based operating systems.
Bowtie 2	[281]	
BWA	[247]	
MAQ	[282]	
SOAP2	[283]	

Table 3.1 Tools for aligning reads to a reference genome. These are some of the many tools built for aligning reads produced from high throughput sequencing. Some have made speed their main purpose whereas others have paid more attention to annotating the files produced (such as mapping quality). Thus a manual review of candidate regions may prove to be crucial especially when dealing with very rare disorders.

Variant calling

There are many tools available for the identification of SNVs, indels, splice-site variants and CNVs present in the query sequence(s). Each variant calling tool has advantages and disadvantages and has made compromises relating to issues such as speed of analysis, annotation and reliability of the output file (Table 3.2). Separating true variation from sequencing artefacts still represents a considerable challenge.

When dealing with very rare disorders, the candidate regions in the output VCF (or BAM) files should be reviewed either by reviewing the QC scores (e.g. base and genotype quality) in the VCF or by visualising the alignments in IGV (or a software of choice) [284]. Performing this step could highlight sequencing errors such as over-coverage (due to greater abundance of capture probes for the region or double capturing due to poorly discriminated probes hybridising to the same region) or under-coverage (due to probes not hybridising because of high variability in the region). For rare Mendelian disorders, since there is going to be a single causal variant it is more important to make sure that the variants in the dataset are reliable. Therefore setting strict parameters for read depth (e.g. $\geq 10x$), base quality score (e.g. ≥ 100) and genotype quality scores (e.g. ≥ 100) initially can eliminate wrong base and genotype calls. This can then be adjusted subsequently (i.e. made less stringent) if no variants with a strong candidacy for causality are found after filtering (also see Best Practices section of GATK documentation for variant analysis).

As mentioned above, there are many tools available for the identification of variants present in the query sequence (see Table 3.2). GATK [249] is one of the most established SNP discovery and genome analysis toolkits, with extensive documentation and helpful forums. It is a structured programming framework which makes use of the programming philosophy of MapReduce to solve the data management challenge of NGS by separating data access patterns from analysis algorithms. GATK is constantly updated and cited, and also has a vibrant forum which is maintained continually.

SAMtools [250] is a variant caller which uses a Bayesian approach and has been used in many WGS and WES projects including the 1000 Genomes Project [208]. SAMtools also offers many additional features such as alignment viewing and conversion to a BAM file. A recent study has compared GATK, SAMtools and Atlas2 and found GATK to perform best in many settings (see reference [285] for details). However all three were highly consistent with an overlapping rate of ~90%. SOAPsnp is another highly used SNP and genotype caller and is part of the reliable SOAP family of bioinformatics tools (<http://soap.genomics.org.cn/>).

Name	References	URL	Notes
GATK	[249]	http://www.broadinstitute.org/gatk/	- Probably the most established genome analysis toolkit - Includes tools such as Unified Genotyper (SNP/genotype caller), Variant filtration (for filtering SNPs) and Variant Recalibrator (for SNP quality scores) - Very well documented with forums - Input: SAM format - Output: VCF format
QCALL	[286]	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	- Theoretically calls ‘high quality’ SNPs even from low-coverage sequencing data - Makes use of linkage disequilibrium information
PyroBayes	[287]	http://bioinformatics.bc.edu/marthlab/wiki/index.php/PyroBayes	- Theoretically makes ‘confident’ base calls even in shallow read coverage for reads produced by Pyrosequencing machines.
SAMTools	[250]	http://samtools.sourceforge.net/	- Computes genotype likelihoods - BCFtools calls SNP and genotypes - Successfully used in many WGS and WES projects such as the 1000 Genomes Project [208]. - Offers additional features such as viewing alignments and conversion of SAM to a BAM format
SOAPSnp	[248]	http://soap.genomics.org.cn/soapsnp.html	- Part of the reliable SOAP family of bioinformatics tools - Well documented website; and cited and used by many [288, 289].
Control-FREEC	[254]	http://bioinfo-out.curie.fr/projects/freec/	- Identifies copy number variations (CNV) between case and controls from sequencing data - R script available for visualising CNVs by chromosome - Input format: BAM
Atlas2	[290]	https://www.hgsc.bcm.edu/software/atlas-2	- Calls SNPs and indels for WES data - Requires BAM file as input - Output: VCF format

Table 3.2 Tools for identifying variation from a reference genome using NGS reads. GATK, SOAPSnp and SAMTools have constantly been cited in large genetic association projects indicating their ease of use, reliability and functionality. However, this is also helped by the fact that they have additional features. There are other tools such as Beagle [291], IMPUTE2 [292] and MaCH [293] which have modules for SNP and genotype calling but are mostly used for their main purpose such as imputation and haplotype phasing

Additional checks of autozygosity

For data obtained from consanguineous families, confirming expected autozygosity (i.e. homozygous for alleles inherited from a common ancestor) would be an additional check worth carrying out. If the individual is the offspring of first cousins, then the level of autozygosity would be approximately 6.25% ($F=0.0625$); and 12.5% ($F=0.125$) for offspring of double first cousins (or uncle-niece unions, see section 9.7 for a depiction of these unions). These values will be higher in endogamous populations (e.g. for offspring of first cousins: 6.25% plus autozygosity brought about due to endogamy - see Figure 1.12 for an example). Autozygosity could be checked by inspecting long runs of homozygosity (LROH) for each individual by using tools such as Plink (for SNP chip data) [294], EXCLUDEAR (for SNP chip data) [295], AgilentVariantMapper (for WES data) [296] and AutoSNPa (for SNP chip data) [297] and dividing total autozygous regions by total length of autosomes in the human genome (can be obtained from http://www.ensembl.org/Homo_sapiens/Location/Genome). AutoZplotter that we (but mostly Dr. Tom Gaunt at the University of Bristol) developed takes VCF files as input, enabling easy and reliable visualisation and analysis of LROH for any type of data (WGS, WES or SNP chip). The code (written in the Python programming language) can also be adapted relatively easily for use in analyses of other species.

3.3.2. Stage 2 – Filtering/Ranking of Variants

Once the quality control process is complete and VCF files are deemed 'analysis ready', the approach taken will depend on the type of disorder analysed. For rare Mendelian disorders, many filtering and/or ranking steps can be taken to reduce the thousands of variants to a few strong candidates. Screening previously identified genes for causal variants is a good starting point. Carrying out this simple check will allow the identification of the causal variant even from a single proband thus potentially saving considerable time, effort and funding. If no previously identified variant is found in the proband analysed, there are several steps which can be taken to identify novel mutations.

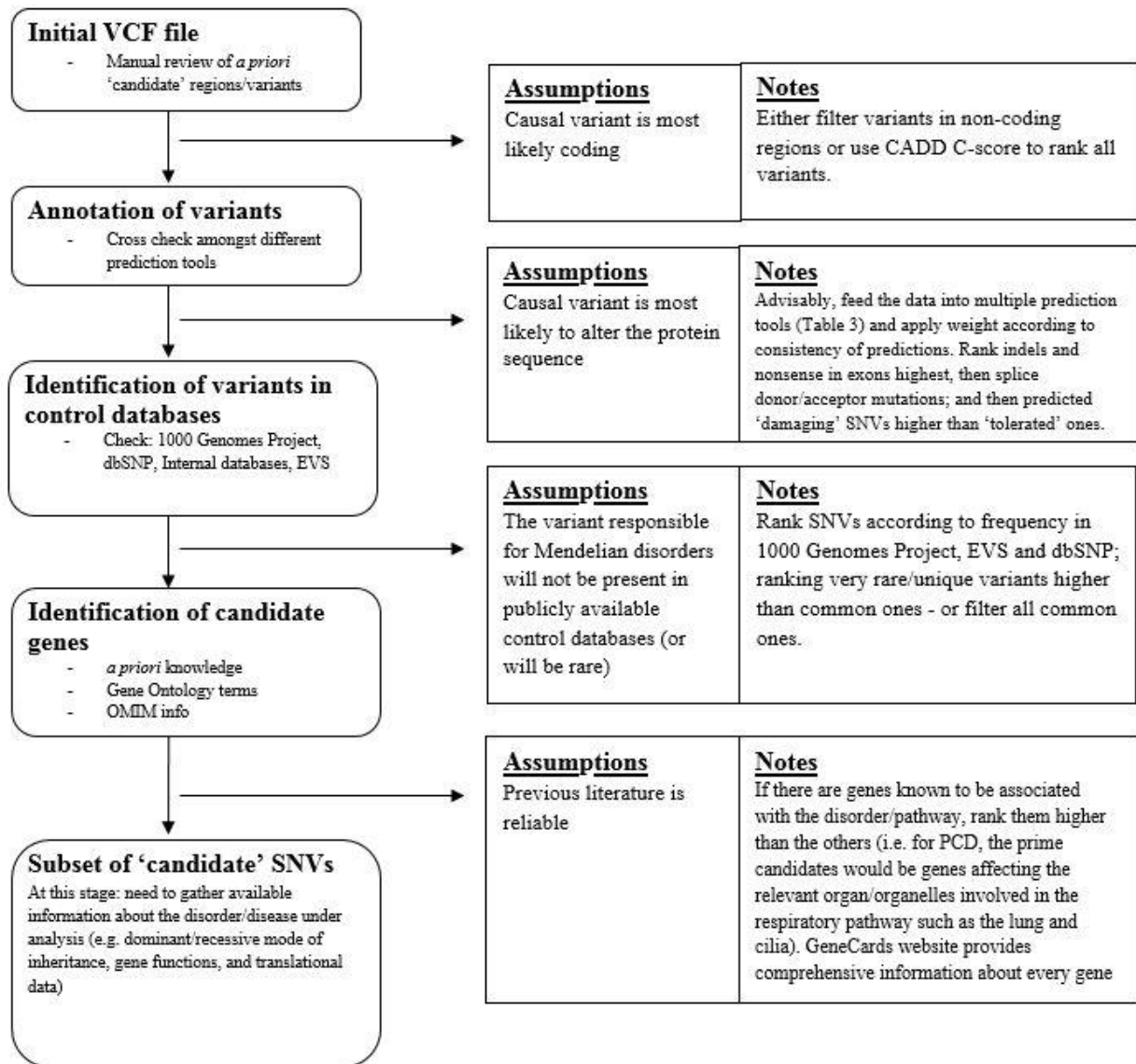


Figure 3.2 Post-VCF file procedures (example for sequencing data). Every step can be automated through the use of pipelines and bioinformatics tools. Whilst performing the steps listed above, one must always bear in mind the assumptions behind the procedures. Ranking of rare SNVs would be advised over filtering as it allows the researcher to observe all variants as a continuum from most likely to least likely.

Using prior information to rank/filter variants

Locus specific databases (see <http://www.hgvs.org/dblist/dblist.html> for a comprehensive list) and 'whole-genome' mutation databases such as HGMD [298], ClinVar [299], LOVD [300] and OMIM [52] are very informative resources for the screening of previously reported variants. Finding no previously identified variants indicates a novel variant in the proband(s) analysed. For rare Mendelian disorders, the search for the 'causal' variant can begin by removal of known neutral and/or

common variants (e.g. $\geq 0.1\%$) as this would provide a smaller subset of potentially causal variants. This is a pragmatic choice as Mendelian disease causal variants are likely to be very rare in the population or unique to the proband. If the latter is true, the variant should be absent from public databases. For this process to be thorough, an automated annotation tool such as Ensembl VEP or ANNOVAR can be used (see reference [301] for a review on the caveats of using these consequence predictors). Ensembl VEP enables incorporation of allele frequency (labelled as GMAF, global minor allele frequency) information from the EVS and the 1000 Genomes Project (see section 9.4 for details).

Using effect prediction algorithms to rank/filter variants

Following on from the previous section, ranking this subset of (rare) variants based on consequence (e.g. stop gains would rank higher than missense) and scores derived from mutation prediction tools (e.g. 'probably damaging' variants would rank higher than 'possibly damaging' according to PolyPhen-2 prediction) would enable assessment of the predicted impact of all rare mutations. It is important to understand what is assumed at each filtering/ranking stage; and comments are included about each assumption and their caveats in Figure 3.2.

For individuals of European ancestry, a VCF file will have between eighty and ninety thousand variants for WES (more for individuals with African ancestry [212]); and approx. a tenth will be variants with 'predicted high impact' (also known as Φ variants i.e. rare nonsense, missense, splice-site acceptor or donor variants, exonic indels, start losses [80]). There are many algorithms which predict the functional effect of these variants (Table 3.3). A large proportion of these algorithms utilize sequence conservation within a multiple sequence alignment of homologous sequences to identify intolerant substitutions, e.g. a substitution falling within a conserved region of the alignment is less likely to be tolerated than a substitution falling within a diverse region of the alignment (see reference [302] for a review). A handful of these algorithms also utilize structural properties, such as the protein secondary structure and solvent accessible surface area, in order to improve

performance. Well known examples of a sequence-based and structure-based algorithm are SIFT [243] and PolyPhen [244] respectively. Newer software such as FATHMM [242] and MutPred [303], which use state-of-the-art hidden Markov models and machine learning paradigms, are worth utilising for their performance. There are also several tools (called ‘meta-predictors’) such as Condel-2 [257] which combine the output of several prediction tools (e.g. FATHMM and Mutation Assessor) to produce a ‘consensus deleteriousness’ score. Although SIFT and PolyPhen are highly cited tools, comparative analyses carried out by Thusberg *et al* and Shihab *et al* found several tools perform better when the VariBench dataset is used as a benchmark [242, 255]. For predicting the effects of non-coding variants, FATHMM-MKL [270], GWAVA [271] and/or CADD [269] should be used. Also Human Splice Finder (latest: v3.0) can be used for intronic variants which predicts whether splicing is affected by the variant or not [304]. Many of these tools can be incorporated into the analyses through the Ensembl website (<http://www.ensembl.org/info/docs/tools/vep/index.html>) where VCF files are annotated [305].

These prediction algorithms are – as their name suggests – only there to make predictions about whether a variant is expected to be functionally disruptive or not. Thus their main purpose is to enable researchers to rank certain variants higher than others in order for them to be studied in a systematic way. Thus they do not ‘prove’ anything about the causality of the variant. The variants predicted ‘deleterious’ still require following up through replication and/or functional studies. Also disagreements amongst different tools can be observed which can lead to different interpretations about the evolutionary history of the variant (e.g. same function conserved throughout different species or a recently acquired function). Users of prediction algorithms should be aware of how these algorithms derive their predictions and then decide whether the tool can be generalized to their datasets. For example, those interested in somatic mutations should choose cancer-specific algorithms e.g. FATHMM-Cancer [306] and SPF-Cancer [307], given that germline variant prediction algorithms are incapable of discriminating between cancer driver mutations and other germline mutations.

Name	Reference	MCC	Comments
*SIFT	[308, 309]	0.30 (unweighted)	Highly cited with many projects using and citing it since 2001. Uses available evolutionary information and is continually updated. Easy to use through VEP. Provides two classifications: 'Deleterious' and 'Tolerated'.
*PolyPhen-2	[244]	0.43	Provides a high quality multiple sequence alignment pipeline and is optimized for high-throughput analysis of NGS data. Cited and used by many projects of different types. Easy to use through VEP. Provides three classifications: 'Probably Damaging', 'Possibly Damaging' and 'Benign'.
*FATHMM	[242]	0.72	A highly performing prediction tool. Clear examples are available on the website. Offers flexibility to the user for weighted (trained using inherited disease causing mutations) and unweighted (conservation-based approach) predictions. Also offers protein domain-phenotype association information.
GERP++ (and GERP)	[310-312]	N/A	Determines constrained elements within the human genome; therefore variants in them are likely to induce functional changes. Can provide unique details about the candidate variant(s).
PhyloP	[313]	N/A	Helps detect non-neutral substitutions. Similar aim with GERP
CADD	[269]	-	Provides annotation and scores for all variants in the genome considering a wide range of biological features
GWAVA	[271]	-	Provides predictions for the non-coding part of the genome.
*SNAP	[314]	0.47	Predicts the effects of non-synonymous polymorphisms. Cited and used many times; and should be used to check whether the predicted effect is matched by the putative causal variant. However it was labelled 'too slow' for high throughput analyses by [255].
PupaSuite	[315]	-	Identifies functional SNPs using the SNPeff [316] database and evolutionary information.
Mutation Assessor-2	[317]	-	Predicts the impact of protein mutations. User friendly website and accepts many formats.
*PANTHER	[318, 319]	0.53 (unweighted)	Predicts the effect of amino acid change based on protein evolutionary relationships. It provides a number ranging from 0 (neutral) to -10 (most likely deleterious) and allows the user to decide on the "deleteriousness" threshold. It is constantly updated making it a very reliable tool.
Condel-2	[257]	-	Combines FATHMM and Mutation Assessor (as of version 2) in order to improve prediction. It theoretically outperforms the tools it is using in comparison to when the tools are used individually.
*MutPred	[303]	0.63	Predicts whether a missense mutation is going to be harmful or not based on a variety of features such as sequence conservation, protein structure and functional annotations. Praised in recent comparative study by [255].
*SNPs&GO	[320]	0.65	Reported to have performed best amongst many prediction tools in [255]. Provides two classifications: 'Disease related' and 'neutral'.
Human Splicing Finder	[304]	N/A	Predicts the effect of non-coding variants in terms of alteration of splicing. Useful for compound heterozygotes if one allele is intronic.
Others	[321], [322], [323], [324]	0.19 0.43 0.40 -	*nsSNPAnalyzer (requires 3D structure coordinates), *PhD SNP, *Polyphen (not supported any more), PMUT

Table 3.3 Tools for predicting variant effects: Identifying neutral and pathogenic mutations. Many methods have been developed to predict the effect of missense mutations. Many of the tools listed above use different features and datasets to predict these effects; thus once the decision is made about which tool to use, the theory behind the predictions should always be kept in mind. Tools such as Condel-2 combine several of these tools to determine a consensus score which theoretically results in higher accuracy when compared to the individual tools. *Comprehensive information about the prediction tool including accuracy, specificity and sensitivity available in [255] and [242]. N/A: not applicable. MCC: Matthew's Correlation Coefficient. MCCs obtained from reference [242].

Further filtering/ranking

With current knowledge, there are approx. fifty synonymous mutations with proven causality – complex traits and Mendelian disorders combined [325]. This is a very small proportion when compared to the thousands of published clinically relevant non-synonymous (i.e. missense and nonsense) mutations. Therefore, when filtering variants for rare monogenic disorders, not taking non-coding variants and synonymous variants into account in the initial stages is a pragmatic choice. If ranking is preferred, then tools such as SilVA [326] which ranks all synonymous variants and CADD [269] which ranks all variants (including synonymous variants) in the VCF files should be used.

Highly penetrant (Mendelian or common-complex) disease causal variants are expected to be very rare, therefore most of them should not appear in publicly available datasets. However filtering all variants present in dbSNP which is common practice, should not be carried out as amplification and/or sequencing errors as well as potentially causal variants are known to make their way into this database (see references [87, 239] for details). Thus use of a MAF threshold (e.g. $\leq 0.1\%$ in 1000 Genomes and/or EVS) is a wiser choice in contrast to using ‘absence in dbSNP’ as a filter. Upon completion of these steps, a smaller subset of variants with strong candidacy will remain for further follow up to determine causality.

Another initially pragmatic choice is to filter out all the annotations except for the ‘canonical’ transcripts (i.e. longest transcript of a gene – if several exist) as this can reduce the amount of variants present in the Ensembl VEP (or ANNOVAR) annotated files considerably (~5x fold). However, this can be a problem for genes where the canonical transcript does not contain all the exons present within the gene – as a mutation which falls in an exon which is not present in the canonical transcript will not be observed in the filtered file (coded ‘CANONICAL’ in Ensembl VEP annotated variants).

As many online tools are expected to keep logs of the processes undergoing in their servers, to protect the confidentiality of genetic information, downloading a local version of the chosen tools (or the VEP cache from the Ensembl website) is recommended. VEP also enables the incorporation of many other annotations (e.g. conservation scores, genomic positions of variants present in 'Public' version of HGMD, PubMed ID of citing publications), which will make the screening and filtering steps more manageable.

3.3.3. Stage 3 - Building evidence for causality

Figure 3.3 (caption of figure on subsequent page) suggests an example route to take to help differentiate causal variant(s) from non-causal ones for Mendelian disorders. At this stage one must gather all information that is available about the disorder and use them to determine which inheritance pattern fits the data and which complications may exist (e.g. the possibility of compound heterozygotes in disorders which show allelic heterogeneity). Figure 3.4 can be used to compare and contrast between the routes taken when analysing Mendelian (Figure 3.3) and common-complex disorders.

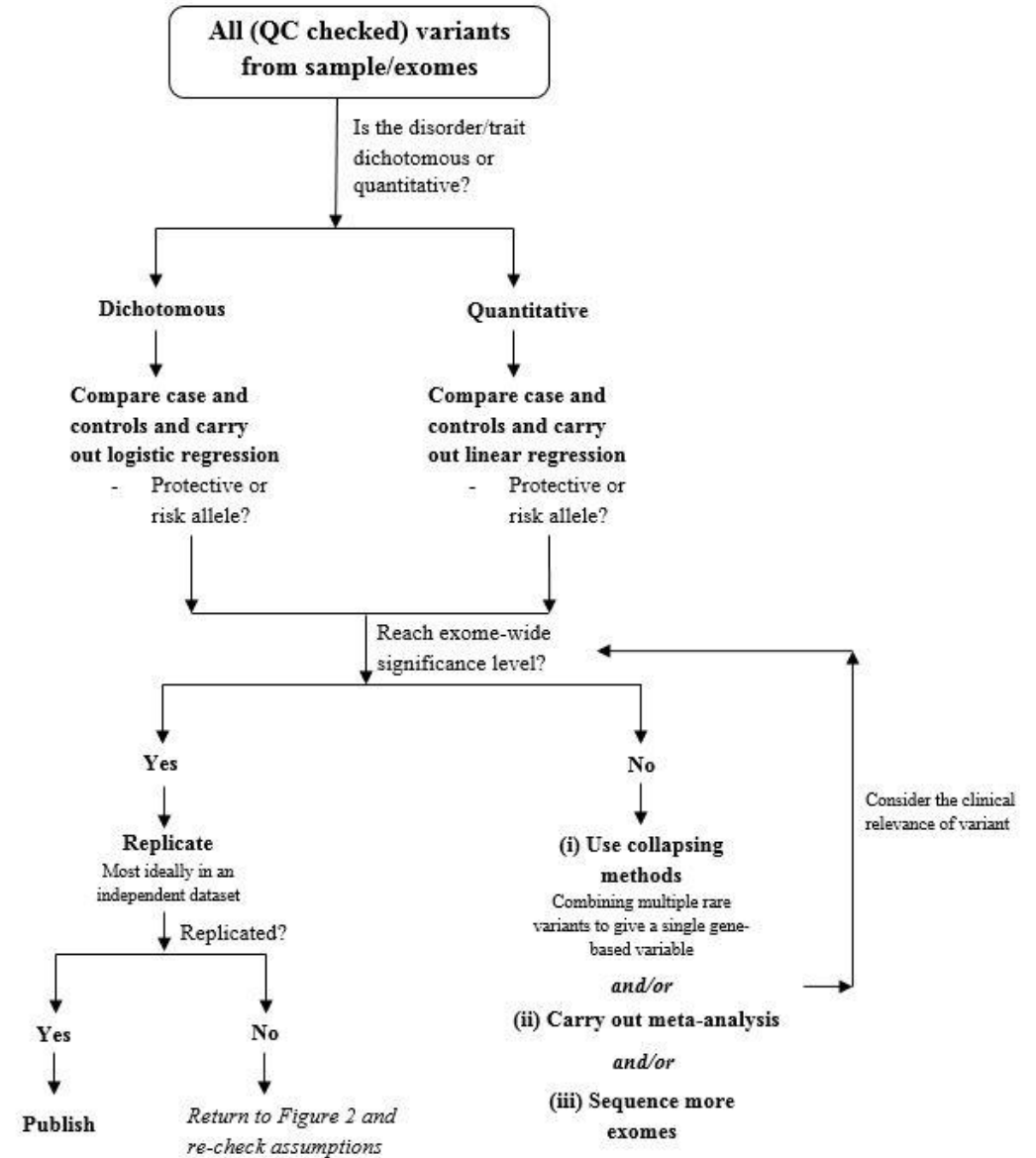
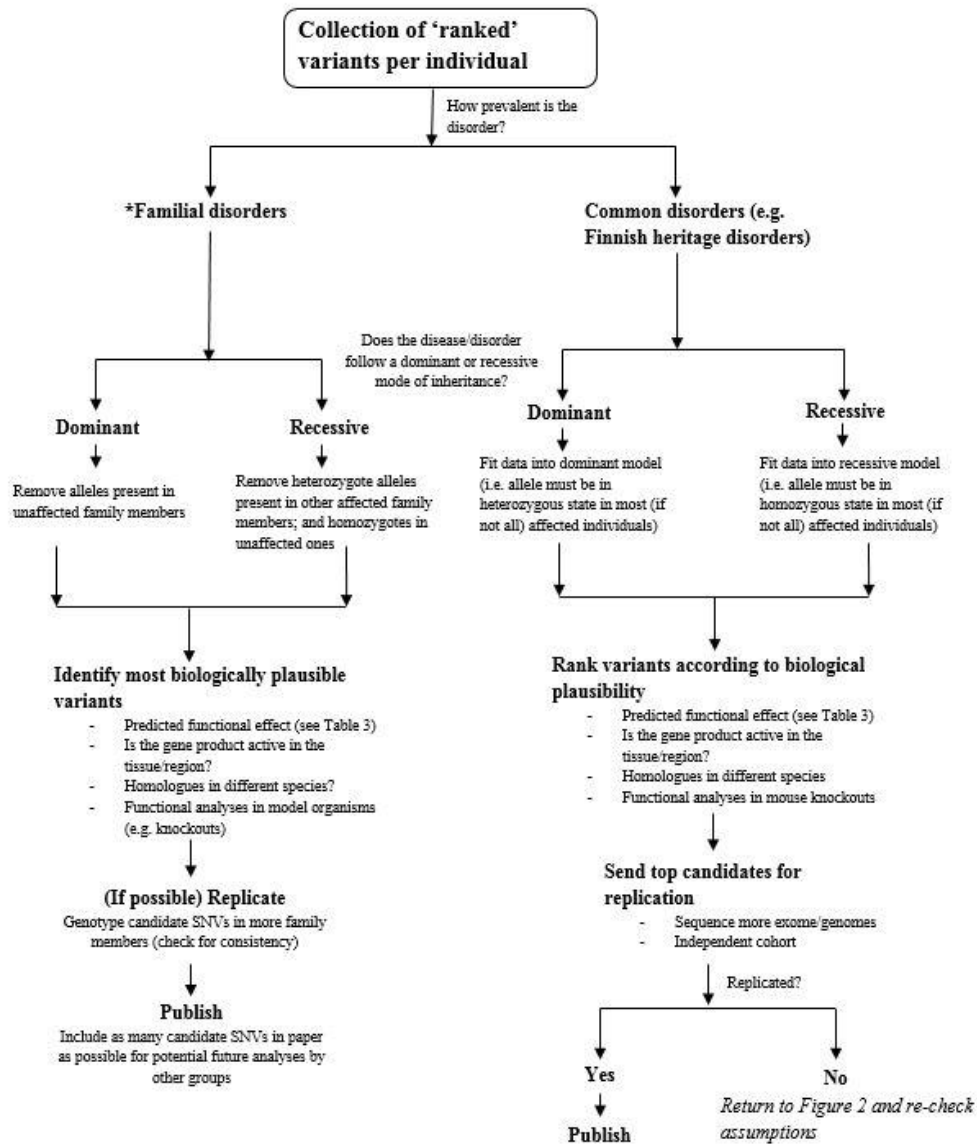


Figure 3.3 (left of dashed line) Finding ‘the one’ in Mendelian Disorders: Searching for the causal variant (WES example). After potentially causal variants are identified, one must put into practice what past literature suggests about the disorder and make certain decisions about which path to follow in Figure 3.5. Familial (very rare) disorders are more likely to be following a recessive mode of inheritance, thus family data is crucial (to rule out *de novo* mutations). Also it is crucial to include as many family members as possible. For common Mendelian disorders, if the disorder is following a recessive inheritance model, the possibility of the existence of compound heterozygotes should be taken into account when fitting the data into a recessive model. Finally, functional post-analysis of candidate variant(s), especially in mouse knockouts, can be crucial.

*If a consanguineous family, identify regions where there are long runs of homozygosity (LROH) for each individual; and amongst these regions, the ones which are shared by the affected and not by the unaffected.

Figure 3.4 (right) Finding ‘the lot’ in Complex disorders: Searching for causal variants (WES example). The standard procedure is to compare cases with controls and detect whether there are any significant differences in the allele frequencies of each variant. The statistical power of this approach is going to predominantly depend on sample size and penetrance of the causal variant. Covariates should be identified and population stratification should be controlled for in the regression models. The clinical significance of the variant must also be taken into account especially when searching for variants with very low effect sizes. One must consider whether it is worth sequencing more exomes in order to reach exome wide significance for the identification of a variant which does not have any considerable effect on patients’ health.

Public data as a source of evidence

Having a candidate gene list based on previously published literature (e.g. by using OMIM or a disease/pathway specific databases such as the Ciliome database [259]) and knowledge about the biology of the disorder (e.g. biological pathways) is useful. Software such as STRING and KEGG (manually curated) predicts protein-protein interactions using a variety of sources [258, 327]. SNPs3D has a user friendly interface which is designed to suggest candidates for different disorders [328]. UCSC Gene Sorter (accessible from <https://genome.ucsc.edu/>) is another useful tool for collating a candidate gene list as it groups genes according to several features such as protein homology, co-expression and gene ontology (GO) similarity. Uniprot's (<http://www.uniprot.org/>) Blast and Align functions can provide essential information about the crucial role a certain residue plays within a protein if it is highly conserved throughout many species. This is especially important for SNVs where the SNV loci itself should be causal (e.g. missense mutations. Excludes nonsense mutations as they truncate the gene product, thus the deleted segment of the protein requires further follow-up to prove causality, not just the loci where the mutation occurred as in other SNVs).

An example of the filtering process for an autosomal recessive disorder such as PCD is depicted in Figure 3.5. If several variants pass the filtering steps, information about the relevant genes should be gathered using databases such as GeneCards (www.genecards.org/) and NCBI Gene (www.ncbi.nlm.nih.gov/gene) for functional information, GEO Profiles (www.ncbi.nlm.nih.gov/geoprofiles) and Unigene (www.ncbi.nlm.nih.gov/unigene) for translational data about the gene's product; and if available, one can check if a homologue is present in different species using databases such as HomoloGene (www.ncbi.nlm.nih.gov/homologene) and whether a similar phenotype is observed in model organisms. For example, if the disorder affects the cerebral cortex but the gene product is only active in the tissues located in the foot, then one cannot make a good argument about the identified variant in the respective gene as being 'causal'.

There are many complications that may arise depending on the disorder such as genetic (locus) heterogeneity [329], allelic heterogeneity [330] and incomplete penetrance [331]. Therefore gathering as many cases from the same family is helpful. However for very rare Mendelian disorders this may not be possible, thus it is important to seek other lines of evidence for causality (e.g. animal models, molecular analyses).

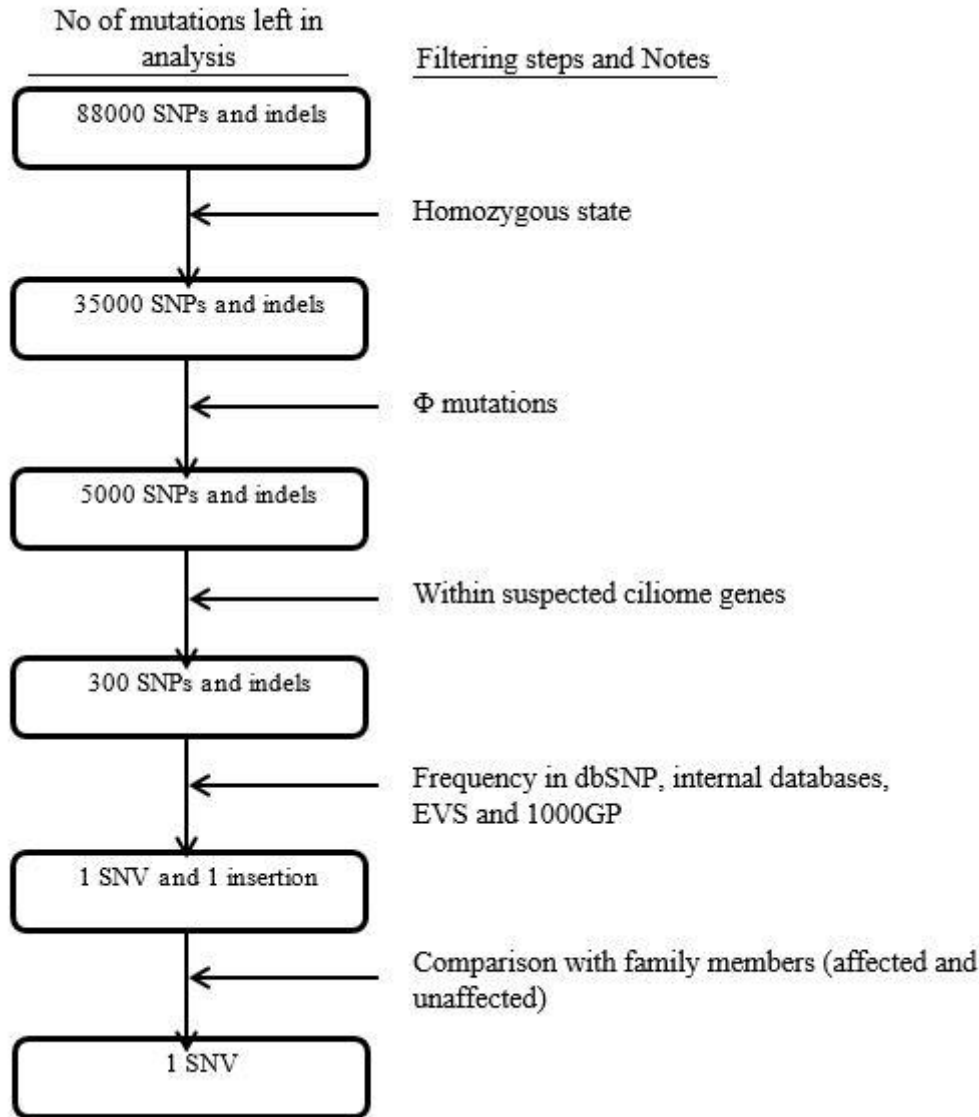


Figure 3.5 Filtering steps applied to all mutations in the exome (Primary ciliary dyskinesia example). After all the filtering steps in the above figure are applied, the total will be reduced to a single candidate. The numbers here are for illustration purposes only (adapted from reference [80]). Homozygosity step is added as PCD is an autosomal recessive disorder. Φ mutations are ‘predicted high impact’ mutations as proposed by Alsaadi and Erzurumluoglu *et al* [80] (see PHI_SO_terms.txt in Section 9.8).

Mapping causal loci within families

For rare Mendelian disorders, familial information can be crucial. The availability of an extended pedigree can be very informative in mapping which variant(s) fits the mode of inheritance in the case(s) and not in the unaffected members of the family (e.g. for autosomal recessive mutations, confirming heterozygosity in the parents is a must). This will provide linkage data where its importance is best displayed by Sobreira *et al* where WES data from a single proband was sufficient in discovering the causal variants in two different families [332]. Where available, previously published linkage data (i.e. associating a chromosomal region to a Mendelian disorder) should also be made use of.

Traditionally, a LOD score of 3 (Prob. = 1/1000) is required for a variant/region to be accepted as causal. Reaching this threshold requires many large families with many affected individuals. However this is not feasible for most highly penetrant disease causal variants (which are very rare by nature) and other lines of evidence such as animal knockouts, molecular studies and local sequence alignments (e.g. by using UniProt as mentioned above) are required to make a case for the causality of variants, especially mutations which are not stop gains (e.g. missense).

As mentioned previously, understanding the characteristics of a Mendelian disorder is important. If the disorder is categorised as ‘familial’ (i.e. occurs more in families than by chance alone), which are usually very rare by nature, then availability of familial data becomes crucial – as unaffected members of the family are going to be the main source of information when determining neutral alleles. Any homozygous (and rare) stop gains, splice-site acceptor/donor variants and start losses in previously identified genes would be prime candidates.

Approach taken in families is different from the approaches taken when analysing common Mendelian disorders using unrelated individuals. For common Mendelian disorders (e.g. Finnish Heritage disorders [333-335]), fitting the dataset into a recessive inheritance model requires most (if not all) affected individuals to have two copies of the disease allele, enabling the identification of founder mutations as

they will be overrepresented in the cases. These variants will be homozygous through endogamy and not consanguinity.

Autozygosity mapping

For consanguineous subjects, the causal mutation usually lies within an autozygous region (characterised by long regions of homozygosity, LRoH, which are generally >5Mb, see [336]), thus checking whether any candidate genes overlaps with an LRoH can narrow region(s) of interest. There are several tools which can identify LRoHs such as Plink, AutoSNPa and AgilentVariantMapper. With this thesis, I have made available a user-friendly python script (AutoZplotter) to plot heterozygosity/homozygosity status of variants in VCF files to allow for manual screening of short autozygous regions as well as LRoHs.

AutoZplotter

There are several software which can detect long runs of homozygosity reliably (>5Mb), however they struggle to identify regions that are shorter. Therefore we developed AutoZplotter which plots homozygosity/heterozygosity state and enables quick visualisation of suspected autozygous regions (requires Xming or another X11 display server). These regions can then be further followed up if any of the identified regions overlaps with a prior candidate gene/region. The input format of AutoZplotter is VCF, thus it will suit any type of genetic data (e.g. SNP array, WES, WGS). AutoZplotter was used for this purpose in a previous study by Alsaadi *et al* [217].

Exceptional cases

There can always be exceptional cases (in consanguineous families also) such as compound heterozygotes (i.e. individuals carrying different variants in the two copies of the same gene). This would require haplotype phasing and the confirmation of variant status (i.e. heterozygosity for one allele and absence of the other) in the parents and the proband(s) by sequencing of PCR amplicons containing

the causal variant (or by genotyping the variant directly). Beagle and HAPI-UR are two widely used haplotype phasing tools for their efficiency and speed [291, 337].

Identifying highly penetrant variants for common-complex disorders

For common complex disorders, identifying causal variants in outbred populations has proven to be a difficult and costly process (Figure 3.4); and these disorders can have many unknowns such as the significance of environmental factors on the disorder (see two examples of differential environmental influence on disease/traits in references [338, 339]) and epistasis [340]. Many of the causal variants may be relatively rare (and almost always in heterozygous state) in the population introducing issues with statistical power. Traditional GWAS do not attempt to analyse them, thus they are largely ignored – leaving a lot of heritability of common complex disorders unexplained. Analysing individuals with extreme phenotypes where the segregation of disease mimics autosomal recessive disorders (e.g. in consanguineous families) can be useful in identifying highly penetrant causal genes/mutations for complex disorders (e.g. obesity and leptin gene mutations [341]). The genetic influence in these individuals is predicted to be higher and the probands are expected to have a single highly penetrant variant in a homozygous state. These highly penetrant mutations can mimic Mendelian disorders causal variants. Therefore similar study designs can be used as stated above (e.g. autozygosity/homozygosity mapping).

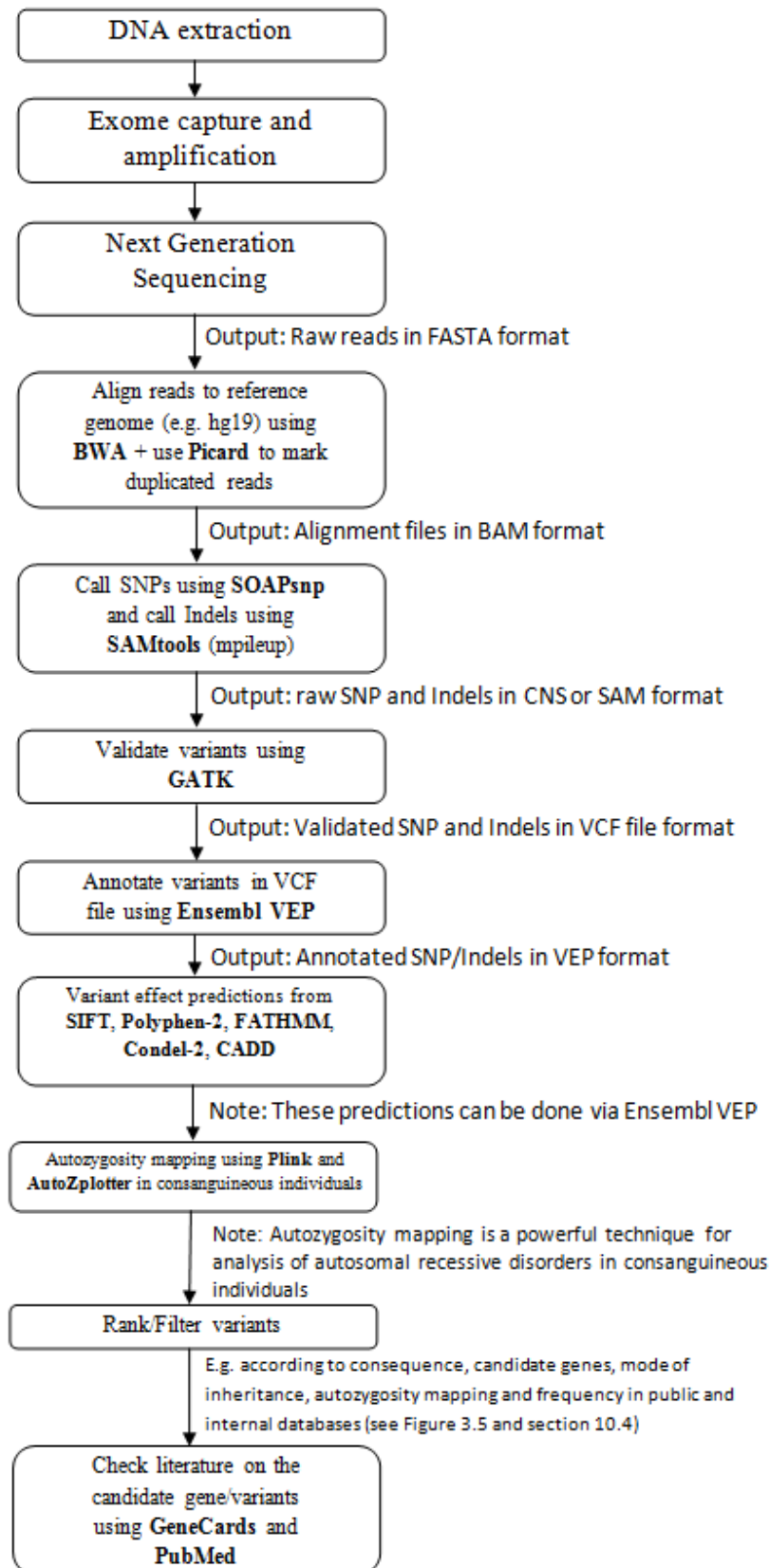


Figure 3.6 Summary of whole analysis process: DNA sample to identification of variant. The tools mentioned here are the ones I prefer to use for a variety of reasons such as documentation, ease of use, performance, multi-platform compatibility and speed. See section 9.4 for examples of parameters/commands to use where applicable.

3.4. Discussion

The NGS era has brought data management problems to traditional geneticists. Many data formats and bioinformatics tools have been developed to tackle this problem. One can easily be lost in the plethora of databases, data formats and tools. “Which tools are out there? How do I use it? What do I do next with the data I have?” are continually asked questions. In this chapter, I aimed to produce a guide to address potential problems in the rapidly changing and ever expanding world of bioinformatics. Figure 3.6 depicts a summary of the analysis process from DNA extraction to finding the causal variant, putting into perspective which file formats are expected at each step and which bioinformatics tools I prefer due to reasons mentioned before. Researchers can then appreciate the stage that they are at and how many other steps are required for completion as well as be guided about what to do at each step.

Whole exome sequencing is the current gold standard in the discovery of highly penetrant disease causal mutations. As knowledge on the non-coding parts of the genome can still be considered to be in its early days, the human exome is still a pragmatic target for many. As approx. 1600 known Mendelian disorders (and ~3500 when suspected ones are included) and most common-complex disorders are still waiting for their molecular basis to be figured out (from omim.org/statistics/entry, true as of 23/06/15), future genetic studies have much to discover. However for these projects to be fruitful, careful planning is needed to make full use of available tools and databases (see Table 3.4).

Finally, with this chapter I have also made AutoZplotter available (input format: VCF), which plots homozygosity/heterozygosity state and enables quick visualisation of suspected autozygous regions. This can be important for shorter autozygous regions where other autozygosity mappers struggle – especially without dense SNP chip array data.

I believe a guide such as this was required in the literature. Although there are overlaps with other reviews, none was structured in this way which can be understood by new comers to the field as well as those who want an update on the ever changing world of bioinformatics. Although just reading this guide alone will not make anyone an expert in the area, thus where necessary I have referred those interested in the details to relevant research articles and comprehensive reviews for.

Material	Notes
'Sufficient' number of high-quality sequencing/genotype data	Amount needed can vary from one proband and a few family members (for very rare Mendelian disorders) to 10000 case and controls (for certain complex disorder/traits)
List of candidate genes	Websites such as OMIM and GHR ; and software such as SNPs3D can be helpful
Identification of variant calling tool	Such as in Table 3.2
Identification of variant effect predictor tool	Such as in Table 3.3; tools usually require conversion of VCF to VEP format (Ensembl website)
Knowledge of human population variation databases	i.e. HapMap, 1000 Genomes Project, EVS, dbSNP, internal databases
Knowledge of databases storing information about genes and their products	i.e. OMIM, Gene (NCBI), GeneCards, Unigene (NCBI), GEO Profiles (NCBI), HomoloGene (NCBI), Mouse knockout databases (such as MGI , TIGM and NC3RS). Search the literature using PubMed and/or Web of Science.

Table 3.4 What is needed for a genetic study? The most important factors when carrying out a genetic association study are (i) the availability of data (ii) expertise and (iii) careful planning

3.5. Conclusions

In this chapter, I presented a generic guide to analysing NGS data. I have structured it in a 'step-by-step' fashion and mentioned the data formats to-be-expected, and the bioinformatics tools that are available for that respective purpose at each stage. I have also made my own judgements about which tool is most suited and stated the reasons behind my choices.

As aforementioned, the information in this chapter has laid the foundations for the analyses carried out in the subsequent chapters.

CHAPTER 4. HUNTING FOR PRIMARY CILIARY DYSKINESIA CAUSAL GENES

In this chapter I introduced Primary Ciliary Dyskinesia (PCD) as a disease, what is known about PCD at current and the findings of this thesis. As PCD requires deeper phenotyping with respect to other Mendelian disorders such as (autosomal recessive form of) Intellectual disability and Papillon-Lèfevre syndrome, the criteria and tools used to diagnose PCD was also included in this chapter.

WES data was obtained (and analysed) from 6 unrelated families. As each family was expected to carry unique PCD causal mutations, each family's results were presented separately. Also as data quality and reliability was crucial in the analysis of rare diseases such as PCD, special attention was paid to the comprehensive presentation of data quality parameters (e.g. read alignment, variant calling) through tables and figures.

4.1. Introduction

Primary ciliary dyskinesia (PCD) belongs to a very diverse group of genetic disorders called the 'ciliopathies', which are characterised by the dysfunction of organelles called cilia (singular: cilium). Cilia are hair-like structures which extend from the cell membrane of nearly every single human cell [342]. Cilia contain a microtubule skeleton (or axoneme) predicted to be assembled from at least 250 evolutionary conserved proteins [259, 343-345]. Cilia carry out many crucial (and some essential) functions such as mucociliary clearance in the respiratory tract [346], establishing the left-right symmetry in the embryo [347], and coordination of ependymal flow in the brain [348].

PCD, also known as ‘immotile cilia syndrome’, is arguably the most prominent genetic abnormality in regards to the respiratory tract and motile cilia [349]. It is a genetically heterogeneous disorder which *usually** follows an autosomal recessive pattern of inheritance and is characterised by sino-pulmonary disease (e.g. respiratory tract infections), laterality defects (e.g. situs inversus[†]) and male infertility [349, 350]. A consensus on the true prevalence of PCD has not been reached in the past due to difficulties with reaching a conclusive diagnosis [351]; and things have not changed in this regard with current estimates ranging between (and within) populations from 1 in 4000 to 1 in 40000 with figures between 1 in 10 thousand and 1 in 20 thousand being the most prevalent ones [351-358]. However, the prevalence of the disorder is noticeably higher in consanguineous and/or isolated populations [359-363].

The basic structure of respiratory cilia is known as the 9+2 structure where nine doublet microtubules surround a central pair of singlet microtubules (Figure 4.1) [364]. Major structures that are attached to the microtubules are the outer and inner dynein arms, radial spokes and nexin links. These structures are vital for the cilia’s movement. Nine regularly spaced radial spokes form a signal transduction scaffold between the dynein arms and the central microtubule pair in the cilia, flagella and sperm using their ‘head’ and ‘stalk’ combination [274], and are thought to regulate dynein induced movement and cilia/flagella wave formation by determining bend direction and shape of wave [365-367].

In the PCD patients, the respiratory cilia have ultrastructural abnormalities (e.g. missing of crucial parts) and/or do not move in a synchronised fashion, causing the bacteria and dust to remain in the lung and cause downstream phenotypic effects such as chronic pulmonary infections.

* There have been instances where it is reported to follow different types of inheritance patterns such as autosomal dominant

[†] A common congenital abnormality characterized by lateral transposition of the viscera (i.e. internal organs in the main cavities of the body, e.g. thoracic, abdominal)

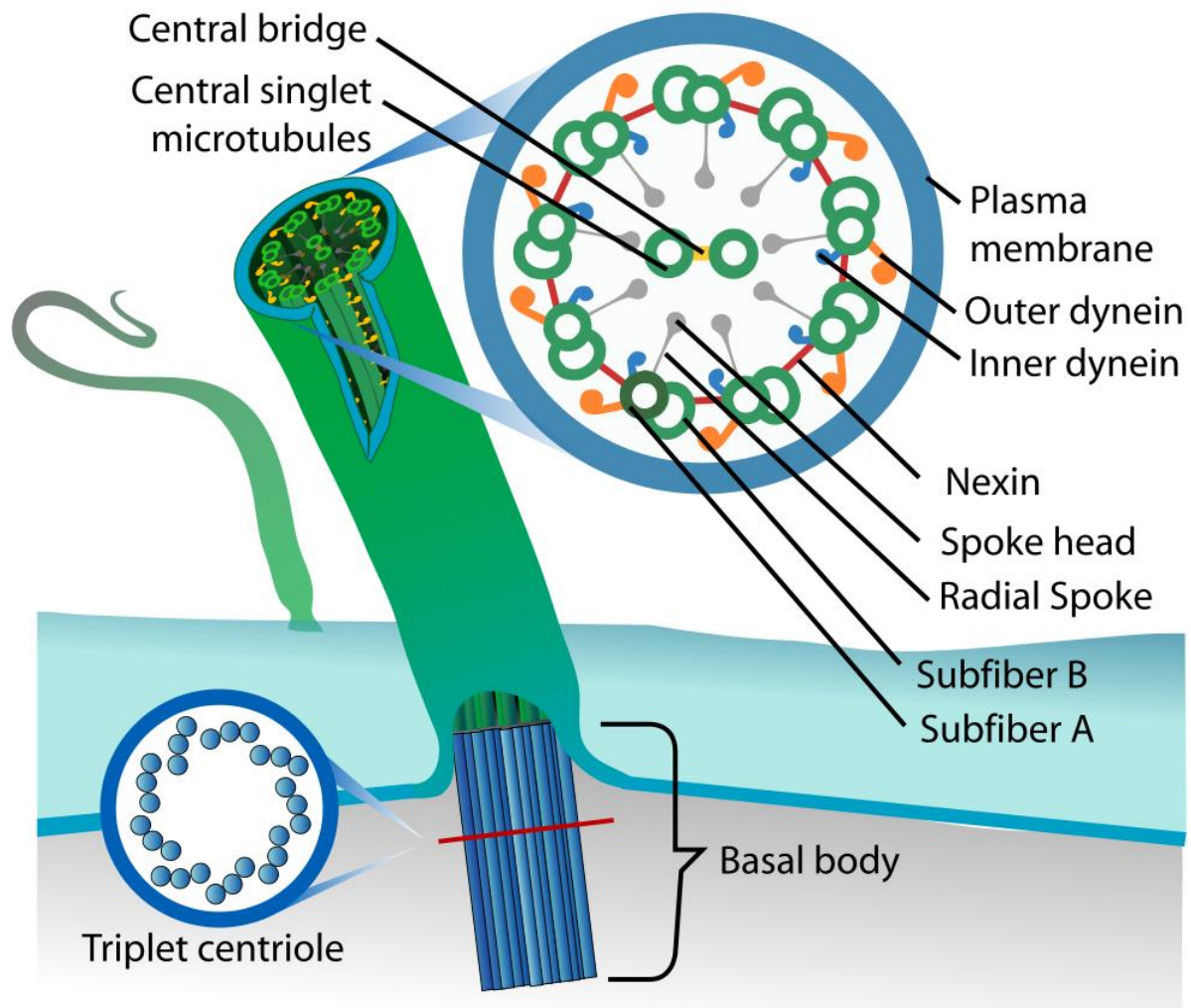


Figure 4.1 Cross-sectional depiction of respiratory cilia with labels for each sub-structure. Image reproduced under the Wikimedia Commons Licence, URL: http://creationwiki.org/pool/images/1/1a/Cilium_anatomy.png.

4.1.1. Diagnosis of PCD in Clinical Settings

As aforementioned, diagnosing PCD requires deep phenotyping as it can easily be misdiagnosed due to the overlapping symptomatology with other respiratory diseases – inherited and acquired.

Besides the chronic respiratory infections due to poor mucociliary clearance and therefore retention of bacteria, low levels of nasal nitric oxide (<10) is also a good indicator of PCD*. However, PCD is usually confirmed by analysing (many) cross-sections and beating patterns of respiratory cilia using electron microscopy (EM) and high definition video cameras.

*Although the reasons behind this observation are not fully understood

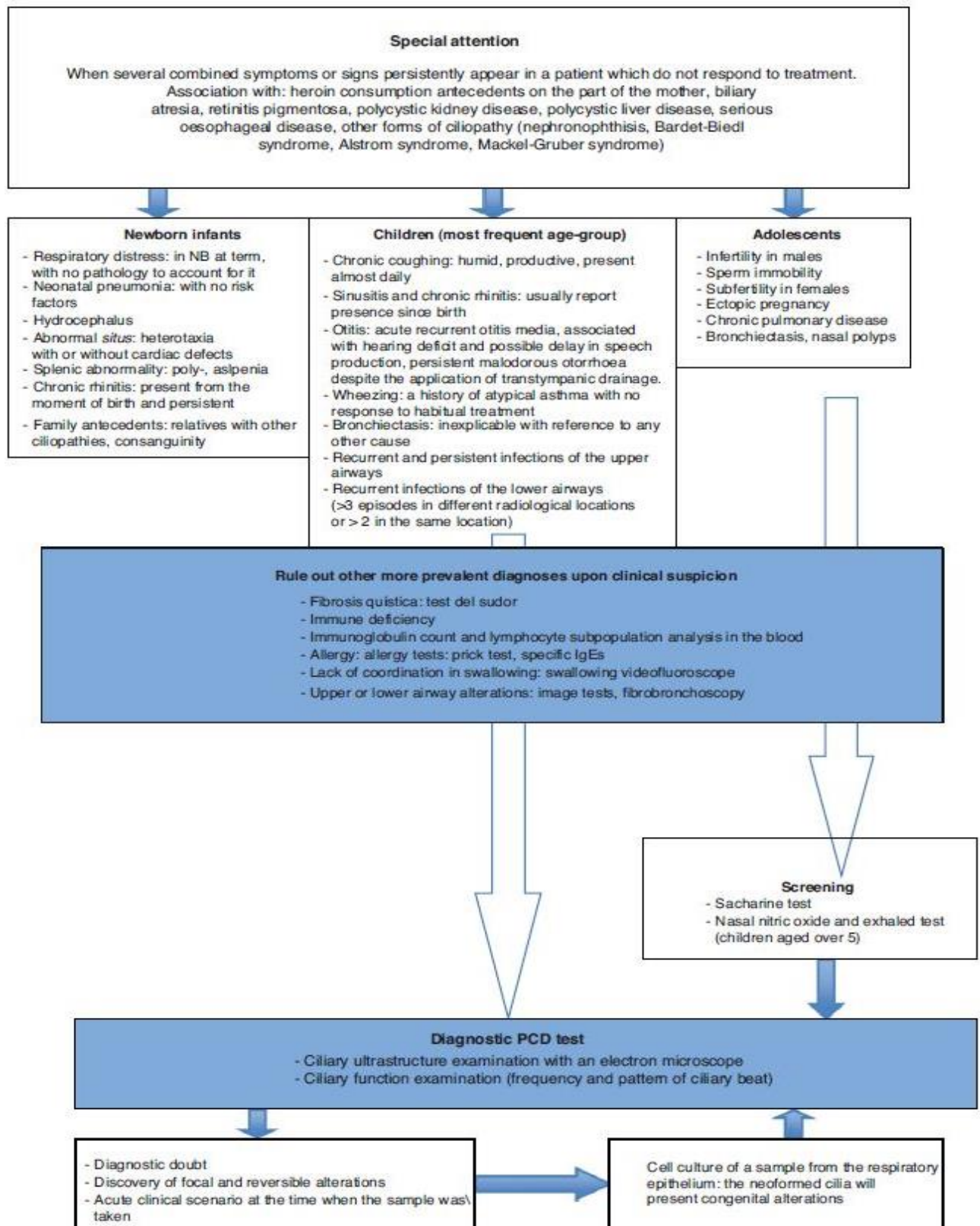


Figure 4.2 Workflow for diagnosing PCD as suggested by Busquets *et al* [368]. Distinguishing PCD from other related conditions is crucial, as misclassification can lead to wrong conclusions in subsequent analyses.

Busquets *et al* have produced a comprehensive workflow which can be followed to dissect PCD patients from individuals with other respiratory diseases (Figure 4.2).

4.1.2. Genetic Aetiology of PCD

As aforementioned in section 4.1, the structure of respiratory cilia is complex. This complexity is mirrored in the genetic aetiology of ciliopathies, and PCD is no different. Unearthing the molecular cause of PCD has therefore been relatively difficult using old technologies such as genome-wide SNP arrays and combining them with traditional methods such as homozygosity mapping [369]. Although PCD is an autosomal recessive disorder, it's genetically heterogeneous nature makes it harder for researchers to pinpoint causal mutations/ genes as every case (or family) in the sample can have causal mutations in different genes. This introduces statistical issues therefore a LOD score of 3 becomes hard to reach. This is reflected by the fact that before 2009, only five PCD causal genes were identified and the genetic basis of over 60% of the PCD cases was unknown [274]. Today, especially with the wider use of WES, this figure has fallen to less than 30%, and there are over 25 genes and/or regions that have been identified to be associated with PCD when mutated (Table 4.1). The most common cause of PCD are mutations in the genes which code for the proteins involved in the structure of the (outer and inner) dynein arms [350] – motor proteins which move the cilia microtubules.

As of July 2015, there are 29 genes/regions that have been reported to be associated with human PCD (Table 4.1). Furthermore, two regions on the long arm of chromosome 15 await further analyses to pinpoint the causal gene (written in **bold** in Table 4.1). Also mutations in the *RPGR* (retinitis pigmentosa GTPase regulator; Gene/Locus MIM Number, hereafter MIM No= 312610) and *OFD1* (oral-facial-digital syndrome 1; MIM No= 300170) genes have been shown to cause PCD concurrently with retinitis pigmentosa and mental retardation respectively (in **red** in Table 4.1). At the gene level, the most common cause of PCD is mutations in *DNAH5* (dynein axonemal heavy chain 5; MIM No = 603335), accounting for a figure between 15 to 21% of all PCD cases [370-373]. Another 2 to 9% of all PCD cases can be attributed to mutations in *DNAI1* (dynein axonemal intermediate chain 1; MIM No= 604366) [33, 374-376], 4 to 5% to *DNAAF1* (dynein axonemal assembly factor 1;

also known as *LRRC50*, leucine rich repeat containing protein 50; MIM No= 613190) [370, 377], 2 to 10% to *CCDC39* (coiled coil domain containing protein 39; MIM No= 613798) [370, 378, 379], 1 to 8% to *CCDC40* (coiled coil domain containing protein 40; MIM No= 613799) [370, 379, 380], 6% to *DNAH11* (dynein axonemal heavy chain 11; MIM No= 603339) [370, 381], 2% to *DNAI2* (dynein axonemal intermediate chain 2; MIM No= 605483) [382], 3% to *LRRC6* (leucine rich repeat containing protein 6; MIM No= 614930) [370, 383], and 1 to 2% to *DNAAF2* (dynein axonemal assembly factor 2; also known as *KTU*, kintoun; MIM No= 612517) [370, 384]. The true prevalence of PCD causal mutations in other genes discovered remains unknown for now [370] , albeit presumably, they are going to account for less than 5% of PCD cases when considered individually. Apart from the dynein arms, disruptions in radial spokes, another structural component of the cilia, has been shown to be causal of PCD. *RSPH4A* (radial spoke head protein 4A; MIM No = 612647) and *RSPH9* (radial spoke head protein 9; MIM No = 612648) were the first two genes to be identified in this respect which were shown to be inactivated in individuals suffering from PCD [274, 359]. These genes' products are thought to be active in the 'head' of radial spokes of unaffected individuals, based on *Chlamydomonas reinhardtii* comparative genomic/homology studies [274, 365]. An individual who is a compound heterozygote for mutations in the *NME8* (NME/NM23 family member 8; also known as *TXNDC3*, thioredoxin domain containing protein 3; MIM No = 607421) [385] gene has also been diagnosed with PCD. An explosion in the number of novel human PCD causal genes identified has occurred since May 2011, with homozygous loss-of-function mutations in *CCDC103* (coiled coil domain containing protein 103; MIM No = 614677) [386], *CCDC114* (coiled coil domain containing protein 114; MIM No = 615038) [387, 388], *CCDC164* (coiled coil domain containing protein 164; MIM No = 615288) [389], *HYDIN* (hydrocephalus-inducing; MIM No = 610812) [390], *HEATR2* (HEAT repeat containing protein 2; MIM No = 614864), *DNAL1* (dynein axonemal light chain 1; MIM No = 610062) and *DNAAF3* (dynein axonemal assembly factor 3; MIM No = 614566) shown to be causal of PCD.

A comprehensive review of the literature shows that 4 large regions (named CILD2, CILD4, CILD5, CILD8) which show a strong association with PCD have been identified in linkage studies [361, 391, 392]. However, further studies carried out in 2012 has enabled pinpointing the causal gene in two of these regions; first one being *DNAAF3* [34] in CILD2, then *HYDIN* [390] in CILD5. In 2008, Geremek and his colleagues attempted to find the causal variant/ gene in the CILD8 region that they had identified in 2006, but to no avail [391, 393]. A detailed list of known and implicated human PCD causal genes can be found in section 9.5.2 (Table 10.1).

How many more human PCD causal genes remain is still an unknown; however one thing that is well known is the importance of early diagnosis of PCD, and this will result in well informed genetic counselling being carried out – reducing short and long term morbidity [357]. For the early diagnostics process to occur at an appropriate level, low cost and high throughput screening methods are required which will most probably require the amalgamation of SNP array technologies with MAF data of known PCD causal mutations.

	Gene/Region (Other names)	Disease Phenotypes	Ensembl ID	Variant(s) identified	Reference(s)	Notes/Methods
1	?	PCD	?	?	[394]	5 male children with PCD from affected mother (3 different fathers)
2	<i>CCDC103</i> (CILD17)	PCD & ODA defect	ENSG00000167131	G128fs*25; H154P	[386]	
3	<i>CCDC114</i>	PCD & ODA defect	ENSG00000105479	A248Tfs*52; and 3 Comp Hets	[387, 388]	
4	<i>CCDC164</i>	PCD & nexin-dynein regulatory complex	ENSG00000157856	K686*	[389]	
5	<i>CCDC39</i> (CILD14)	PCD & Axonemal disorganisation & Absent inner dynein arms	ENSG00000145075	E731Dfs*31; T358Qfs*3; S786Ifs*33; 357+1G>C	[378, 379, 395]	~5%
6	<i>CCDC40</i> (CILD15)	PCD & Axonemal disorganisation & Absent inner dynein arms	ENSG00000141519	A83Vfs*82; R942MinsW	[379, 380, 395]	~5%
7	CILD4 (15q13.1-15.1)	PCD	?		[361]	
8	CILD8 (15q24-25)	PCD & Situs inversus	?		[391]	
9	<i>DNAAF1</i> (LRRC50/CILD13)	PCD & Outer/Inner dynein arm defects	ENSG00000154099	E42_K117del; L172R; P451Afs*5	[377, 396]	~5%
10	<i>DNAAF2</i> (C14orf104/KTU/PF13/CILD10)	PCD & Outer/Inner dynein arm defects	ENSG00000165506	S8X; G406Rfs*90	[384]	~1%
11	<i>DNAAF3</i> (CILD2)	PCD & Outer/Inner dynein arm defects	ENSG00000167646	L108P; R136*; V255C	[34]	
12	<i>DNAH11</i> (CILD7)	PCD (Normal ultrastructure)	ENSG00000105877	Y4128*; A4518_A4523delinsQ; Y4128*	[381, 397]	~6%
13	<i>DNAH5</i> (CILD3)	PCD & ODA defect	ENSG00000039139	>40 mutations	[371-373]	~20% of PCD cases
14	<i>DNAI1</i> (CILD1)	PCD & ODA defect	ENSG00000122735	>15 mutations	[374, 375]	~7%

15	<i>DNAI2</i> (CILD9)	PCD & ODA defect	ENSG00000171595	I116Gfs*54; R263*; 346-3T>G (IVS3-3T>G)	[382]	~2%
16	<i>DNAL1</i> (CILD16)	PCD & ODA defect	ENSG00000119661	D150S	[398]	
17	<i>HEATR2</i> (CILD18)	PCD & Outer/Inner dynein arm defects	ENSG00000164818	L795P	[399]	
18	<i>HYDIN</i> (CILD5)	PCD & central pair defects	ENSG00000157423	K307*	[390]	
19	<i>LRRC6</i> (CILD19)	PCD & Outer/Inner dynein arm defects	ENSG00000129295	K200Efs*3; Q192*; E193Rfs*4; A74P; D146H	[383, 399]	~3%
20	<i>NME8</i> (TXNDC3/CILD6)	PCD & ODA defect	ENSG00000086288	(Comp Het) L426* & intronic 271-27C>T	[385]	
21	<i>OFD1</i>	PCD & Mental retardation & Macrocephaly	ENSG00000046651	2122-2125dupAAGA	[400]	
22	<i>RPGR</i>	PCD & Retinitis Pigmentosa & Sensory hearing defects	ENSG00000156313	631_IVS6+9del	[401]	
23	<i>RSPH4A</i> (CILD11)	PCD & Microtubule disorganisation	ENSG00000111834	Q154*; (Comp Het) Q109* & R490*	[274]	
24	<i>RSPH9</i> (CILD12)	PCD & Microtubular pair abnormalities	ENSG00000172426	K268del	[217, 274]	
25	<i>ARMC4</i>	PCD & ODA defect	ENSG00000169126	S892*; E658*	[402]	
26	<i>DYX1C1</i> (DNAAF4)	PCD & Outer/Inner dynein arm defects	ENSG00000256061	T85Rfs4*; E109*; Y128*; V132*; W162*; R270*	[403]	
27	<i>ZMYND10</i> (BLU)	PCD & Outer/Inner dynein arm defects	ENSG00000004838	V16G; L266P; L39P	[404, 405]	
28	<i>SPAG1</i>	PCD & Outer/Inner dynein arm defects	ENSG00000104450	M1?	[406]	

29	<i>C21orf59</i>	PCD & Outer/Inner dynein arm defects	ENSG00000159079	R98*; Y245*; R33W	[407]	
30	<i>RSPH1</i>	PCD & Microtubule disorganisation	ENSG00000160188	W94*; G29*	[408]	

Table 4.1 Currently known human PCD causal/associated genes and/or regions (as of Jun 2015). The full list of all known and potential human PCD causal genes can be found in Table 10.1 in section 9.5.2 (as of July 2015). Table compiled from respective references and review by Kurkowiak *et al* [409]. Genes in red are reported to cause another disease alongside PCD.

4.2. Hypothesis

The hypothesis of the analyses carried out in this chapter was that the PCD symptoms observed in the nine analysed consanguineous individuals are due to the presence of autosomal recessive mutations in autozygous regions (therefore the mutation will be in a homozygous state). Where there were several affected siblings, the same mutation is expected to be in a homozygous state in all.

4.3. Aims and Objectives

The aim of this chapter was to identify PCD causal variants and genes in 6 unrelated families with PCD affected members who are the offspring of consanguineous parents. Where a novel variant was found, the prevalence of the variant was screened in the local population if feasible.

4.4. Methods

Below are the methods used to analyse the six families who participated in this study. The scripts and files used in the analyses can be found in the appendices (section 9.5).

4.4.1. Clinical criteria used for PCD

Since PCD is a phenotypically heterogeneous disorder, there are several routes a clinician may follow to make a diagnosis. The following are the symptoms that are looked for and the tests which are carried before clinically diagnosing an individual with PCD in this thesis:

- 1- Nasal nitric oxide (NNO) levels less than 10
- 2- Electron microscope image of cilia
- 3- Chronic respiratory infections

- 4- *Situs inversus* (if combined with 1-3, then the diagnosis will be Kartagener's syndrome)

These clinical diagnoses were carried out by Prof. Muslim Alsaadi at the King Saud University (Riyadh, Kingdom of Saudi Arabia).

Electron Microscopy

Endoscopic nasal biopsy was taken from the posterior portion of the inferior turbinate. A piece of tissue measuring 2mm in (maximum) diameter was obtained and fixed in 2% buffered glutaraldehyde. Following fixation, the tissue was post-fixed in a buffered solution of osmium tetroxide in order to enhance the contrast.

For ultrastructural examination, the tissue was subsequently plastic embedded and ultrathin sections are cut using a diamond knife. The ultra-thin sections were mounted on a grid and sequentially stained by immersing the grid in solutions of lead citrate and uranyl acetate.

Semi-thin sections were cut at a thickness of 0.5-1 μm and stained with toluidine blue. The semi-thin sections were used to guide the selection of the area to be viewed in ultra-thin sections. The ultra-thin sections were then examined using the JEOL transmission Electron Microscope (EM, model: JEM-1400).

The EM images used in this thesis were produced by Dr. Mohammad Mubarak at the King Saud University.

4.4.2. Participants

All nine subjects had low nasal nitric oxide levels, which is a good predictor of PCD (e.g. measurement of individual 1 from family 6 is 5). In addition to the conventional PCD related phenotypes (e.g. bronchiectasis, bronchial asthma and recurrent chest infections), individual 1 from family 2 and individual 1 from family 4 had *situs inversus* (and therefore Kartagener's syndrome); and rather unusually, individual 1 from family 3 had thrombocytopenia* and tracheoesophageal fistula†. Individual 1

* Low platelet count

† Abnormal connection (fistula) between the oesophagus and the trachea

from family 5 also has a newly-born sibling diagnosed with Kartagener's syndrome (strengthening this diagnosis). The term 'family' is used throughout the thesis to describe a nuclear family consisting of children and parents.

A pedigree for the families could not be provided in this chapter as the information received about the families was incomplete. All information that is known has been presented in their respective sections (e.g. known information about family 1 can be found in section 4.5.2).

4.4.3. Collection/storage of samples and DNA extraction

Peripheral blood samples were collected from all participants in the PCD studies. Details on storage and DNA extraction are as mentioned before and can be found in sections 2.1 and 2.3.1.

4.4.4. DNA sample quality and quantification

As stated in section 2.4.1, the whole-exome sequencing data used in this thesis were produced at the BGI-Tech sequencing centre. Their sample requirements were:

- (i) DNA Sample purity: $OD_{260}/_{280} = 1.8-2.0$
- (ii) DNA sample concentration: $>50\text{ng}/\mu\text{l}$
- (iii) Total DNA sample quantity: $>6\mu\text{g}$

After samples were rested in room temperature, samples 1 and 4 were diluted by adding $10\mu\text{l}$ water; samples 2, 5 and 6 were diluted by adding $20\mu\text{l}$ water, and sample 9 was diluted by adding $25\mu\text{l}$ water. All samples were centrifuged and fully mixed.

$1.5\mu\text{L}$ of DNA solution was then loaded onto an agarose gel (concentration 1%) and electrophoresed for 40 minutes for DNA sample quality check.

4.4.5. Whole-exome sequencing

Data was downloaded from the BGI Hong Kong server (<http://cdts.genomics.hk>) using FTP (other options are listed in 'How To Download Data From CDTs' booklet made available by BGI at the link: <http://cdts.genomics.org.cn/customerSupport/HowToDownloadDataFromCDTS.pdf>). The UNIX command used to download the files was*:

```
wget -r 'ftp://project_name:*password*@cdts.genomics.hk'
```

Quality control on sequencing data

BGI-Tech defined raw reads as 'reads which contain the adapter sequence and/or high content of low quality bases'. These reads were removed before the read alignment process as they could cause mismapping and variant calling issues. BGI-Tech's filtering steps were as follows:

- i) Removal of the adapter reads: An 'adapter' read was defined as a 'contiguous sequence which includes the adapter sequence', and those adapter reads were removed from the raw FASTQ data
- ii) Removal of the low-quality reads: if more than half of bases in a read were low-quality bases, which has been defined as base quality less than or equal to 5, those reads were removed from the raw FASTQ data
- iii) Removal of reads with >10% undetermined bases

After filtering, the remaining reads were called 'clean reads' and used for downstream bioinformatics analysis (e.g. mapping to human reference genome, variant calling). Details on subsequent bioinformatics analyses can be found in section 2.4.2.

* Downloading the complete WES data (in FASTA, BAM and VCF formats) of 9 individuals (total size $\approx$ 90GB) took approximately 5 days to download.

4.4.6. Autozygosity mapping

The bioinformatics tools Plink [294] and AutoZplotter [410] (developed mostly by Dr. Tom Gaunt at the University of Bristol) were used to detect LRoHs (as a proxy for autozygous regions) in the whole-exomes of the participants.

Commands used to run the two tools are as follows:

AutoZplotter

python AutoZplotter.py (required Xming Server opened and X11 Tunnelling enabled in PuTTY on a remote server)*

Output was saved as a png file† and analysed manually for autozygous regions.

Plink

```
plink --file <ped/map> --homozyg --noweb --homozyg-window-kb 1000 --homozyg-window-het 1 --homozyg-group --out <output>
```

```
grep -v -f phom_file.txt output.hom
```

4.4.7. Variant prioritisation procedures

Candidate genes

Initially I created two lists of genes to be reviewed as ‘prime candidates’ in the proband. The first one (hereafter called list 1, see Human_PCD_genes.txt in section 9.5.3) contained the Ensembl IDs of all the known human PCD genes aforementioned. The second (hereafter list 2, see Suspected_Ciliome_genes.txt in section 9.5.3) had the Ensembl IDs of all the genes (except known ones) in the Ciliome database (see Table 10.2, last updated: 24th Dec 2007 [259]) including additional genes which matched the keywords ‘dynein’, ‘radial spoke’, ‘nexin link’ and/or ‘cilia’ in the GeneCards website (www.genecards.org, v3.11) [411].

* Autozplotter.py is available in section 10.8

† Some of the images inserted in this thesis have been edited to ensure anonymity/confidentiality of the participants

The procedure:

- 1- Access Ciliome database at:
http://www.sfu.ca/~leroux/ciliome_database.htm
- 2- Use 'Advanced search' and download all 'Homo sapiens' data
- 3- Copy to MS Excel (removing empty rows); and save as a CSV file
- 4- To grab list of all the genes' Ensembl Ids*:
python **PCD_candidate_genes.py**

Looking for causal variants

After the two candidate gene lists were created, a list of Ensembl VEP's sequence ontology (SO) terms for variants likely to be causal was also created (i.e. predicted high impact variants, called Φ mutations by Alsaadi and Erzurumluoglu *et al* [80]). The latter list can be found in section 9.8 (under PHI_SO_terms.txt). These two lists were used to filter and prioritise certain genes and variants within these genes. The details of the method (and commands) used to analyse the WES data can be found in the appendices (section 9.5.3).

4.4.8. Mutation screening and variant validation**Family 1 and the p.R263* variant in *DNALI1***

PCR was used to amplify a region 235bp long (containing the stop gain) in the family members and was followed on by Sanger sequencing of the amplicons to validate their variant status using a different method other than WES (Table 4.2).

Primer	Base
Forward	5'- AAATGTGAAGCCACTGAGAAGC-3'
Reverse	5'- GCTTCCTTTATCCTTTGGCAG-3'

Table 4.2 The primers used to amplify a 235bp long region containing the p.R263* mutation in the *DNALI1* gene

* The python script can be found in section 10.5.4

Family 2 and the p.R483Q variant in *CCT4*

PCR was used to amplify a region 238bp long (containing the variant) in the affected and the other members of the family (Table 4.3). The PCR amplicons were then sequenced using Sanger sequencing to validate their variant status.

Primer	Base
Forward	5'- AGTGGTATGGAATCCTACTGCG -3'
Reverse	5'- CATTCAACAGATGACTACAAAGC -3'

Table 4.3 The primers used to amplify a 238bp long region containing the p.R483Q mutation in the *CCT4* gene

Family 6, Saudi population and the p.E309* mutation

For family 6, PCR was used to amplify a region 221bp long (containing the stop gain loci) in these individuals (Table 4.4); and these fragments were digested using the *AvrII* enzyme (following manufacturer New England Biolabs UK Ltd, Herts, SG4 0TY's protocol, catalogue no: R0174L, see Table 4.5 for details) and viewed using 96-well microplate array diagonal gel electrophoresis (MADGE) [412] to check for the presence of the p.E309* mutation in the family members and the Saudi population (for details on sample population and ethics statement, see section 2.2). Nucleotide numbering system uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon (Met) as codon 1.

Primer	Base
Forward	5'-AAATGGGAGAAGGCCTAGGATG-3'
Reverse	5'-GAACCAGCTGCAGTACCTAGAG-3'

Table 4.4 Primers used to amplify a 221bp long region containing the p.E309* mutation in the *CCDC151* gene

Enzyme	Cut site	Unaffected	Affected	Size of PCR fragment for unaffected after digestion	Sizes of PCR fragments for affected after digestion
<i>AvrII</i>	5'-C CTAGG-3'	5'-GCC G AGGAG-3'	5'-GCC T AGGAG-3'	221bp	84bp and 137bp

Table 4.5 Digestion of *CCDC151* amplicons using *AvrII* enzyme. The *AvrII* enzyme will digest the PCR amplicons produced using the primers in Table 4.4 where it comes across the sequence CCTAGG (cutting between the two cytosine bases). Since the unaffected individuals did not have the thymine base required for digestion (highlighted in yellow), the PCR amplicons stayed undigested (i.e. 221bp long).

4.4.9. Protein structure modelling

The amino acid sequence of the gene under analysis was copied from UniProt and pasted to the Robetta server for Ginzu prediction (see section 2.4.5 for details on the method) and then for complete structure prediction [260]. Likewise for the mutant version, the amino acid sequence was changed (or truncated) accordingly from the initial amino acid sequence. Depending on the queue at the Robetta server (<http://robetta.bakerlab.org/queue.jsp>), the whole structure modelling process took between a few weeks to a few months. Different domains within the (outputted) protein structures were represented with different colours. The position of the variants (at an amino acid level) were also marked in the outputted structures (e.g. such as in Figure 4.10).

The amino acid sequences submitted for MNS1p, DNAAF3p, CCDC151p and DNALI1p are available in section 9.5.6.

4.5. Results

The results have been divided into 6 sections, one for each family analysed. A detailed explanation for each figure and table were presented for Family 1; and these will not be included for the other families as to avoid repetition (as they will be the same).

4.5.1. DNA sample quality and concentration

BGI-Tech's requirements for WES were stated in section 2.3.1. To meet these standards, several tests were carried out in house and in their labs to ensure high quality WES data. The results are summarised in Table 4.6 and Figure 4.3 below.

Sample No	Lane No	Concentration (ng/μL)	Volume (μL)	Total Mass (μg)	Sample Integrity	Grade by BGI-Tech and Notes
Family 1 Individual 1	1	229	38	8.7	Slightly degraded	Level A
Family 1 Individual 2	2	196	22	4.31	Slightly degraded	Level B
Family 2 Individual 1	3	50.9	62	3.16	Slightly degraded	Level B
Family 2 Individual 2	4	71	46	3.27	Slightly degraded	Level B
Family 2 Individual 3	8	57	61	3.48	Slightly degraded	Level B
Family 3 Individual 1	5	121	33	3.99	Nearly completely degraded	Level D, proposed to resend sample*
Family 4 Individual 1	6	102	33	3.37	Slightly degraded	Level B
Family 5 Individual 1	7	38.2	47	1.8	Slightly degraded	Level C, total mass was too low [±]
Family 6 Individual 1	8	74.2	42	3.12	Slightly degraded	Level B

Table 4.6 DNA sample test results for the 9 individuals used in this study (complementary to Figure 4.3 below). For details on what the grades mean, see section 2.3.1. *See section 9.5.1 for subsequent actions. [±]Sample was doubled.

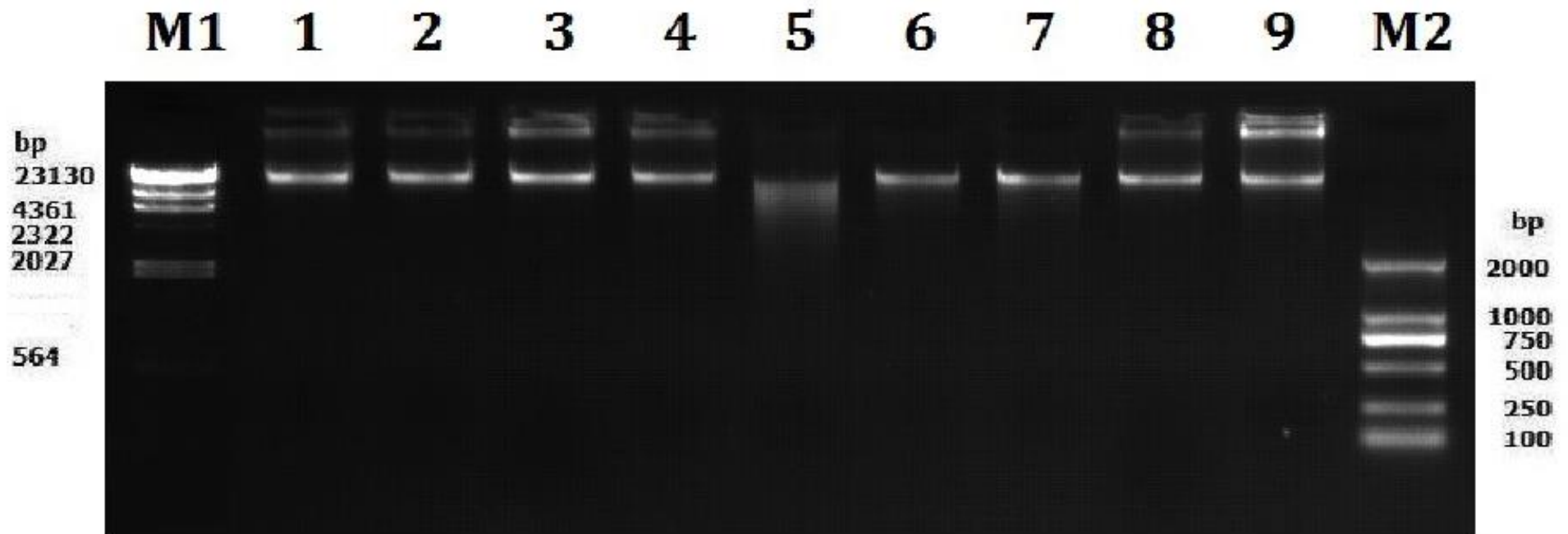


Figure 4.3 Electrophoretogram result for DNA sample integrity. All DNA used in the study were of high quality, except sample 5 where there seems to be some degradation. M1: λ -Hind III digest. M2: D2000 (Tiangen). 1-9: See Table 4.6 above for complementary results.

4.5.2. Family 1

This section will present the results for family 1 and describe the findings:

No of whole-exome sequenced participants: 2

No of other family members with DNA available: 4 (i.e. both parents and two unaffected siblings)

WES data statistics

These results are derived from the WES data produced in section 4.4.5. Since the sequencing strategy used is paired-ends, fq1 represents read 1 and fq2 represents the reads produced via sequencing from other end of DNA fragment. GC(%) means GC content of the reads; and Q20(%) and Q30(%) means the proportion of each read with base qualities more than 20 and 30 respectively. Details can be found in section 4.4.5.

Total captured region for individual 1 was 120,361,800 base pairs (50,743,793 bases on target and 69,618,007 bases near target, the latter being flanking regions within 200bp of exons). Coverage of target (i.e. exons) and flanking regions (e.g. introns, splice sites) was 98.4 % and 95% respectively. The average sequencing depth on target was 60.75 and the fraction of target covered with at least 20 and 10 reads was 78.9% and 88.6% respectively (and >4 read depth = 94.9%). There were a total of 69,186,589 (high quality) reads with a mapping rate of 98.66%.

For individual 2, total captured region was 118,361,446 base pairs (50,599,905 bases on target and 67,761,541 bases near target, the latter being flanking regions within 200bp of exons). Coverage of target (i.e. exons) and flanking regions (e.g. introns, splice sites) was 98.2% and 92.5% respectively. The average sequencing depth on target was 60.75 and the fraction of target covered with at least 20 and 10 reads was 79.4% and 88.6% respectively (and >4 read depth = 94.6%). There were a total of 51,084,667 (high quality) reads with a mapping rate of 99.39%.

Table 4.7 summarises the data produced by the WES platform for Family 1.

Type	Individual 1	Individual 2
Number of Reads	72742728	53684690
Data Size	7274272800	5368469000
N of fq1	38893	28144
N of fq2	29464	21954
GC(%) of fq1	43.46~43.56	44.64~44.74
GC(%) of fq2	43.61~43.71	44.76~44.85
Q20(%) of fq1	96.63~96.82	96.45~96.65
Q20(%) of fq2	95.19~95.36	94.96~95.11
Q30(%) of fq1	89.81~90.28	89.27~89.75
Q30(%) of fq2	87.73~88.14	87.11~87.51

Table 4.7 Family 1 Individual 1 and 2's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values and close to each other between the two individuals.

Read alignment statistics

Below are the read alignment statistics derived from the BWA read alignment tool [247]. Regions 'near target' refers to flanking region within 200bp of target regions. 'Effective reads' corresponds to the number of mapped reads after removal of duplicates. The 'effective bases' means the bases in the effective reads. The term 'target' regions used here refers to genomic regions that the exome array should have covered. 'Reads uniquely mapped' corresponds to the number of aligned reads with a mapping quality ≥ 1 .

The distribution of per-base sequencing depth and cumulative depth distribution in target regions were also plotted. Distribution of per-base sequencing depth approximately followed a Poisson distribution, which shows the exome-capturing target region was evenly sampled. The x-axis denotes sequencing depth, while the y-axis indicates the percentage of total target region under a given sequencing depth. In the plot of cumulative depth distribution in target regions, the x-axis denotes sequencing depth, and the y-axis indicates the fraction of bases that achieves the stated (or above) sequencing depth.

The explanation above will only be stated once and will apply for all tables and figures representing read alignment statistics in all families.

Table 4.8 summarises the quality of alignment of the reads produced by WES platform to the reference genome for family 1.

Exome Capture Statistics	Individual 1	Individual 2
Initial bases on target	51543125	51543125
Initial bases near target	73252334	73252334
Initial bases on or near target	124795459	124795459
Total effective reads	69186589	51084667
Total effective yield(Mb)	6801.89	5024.12
Average read length(bp)	98.31	98.35
Effective sequences on target(Mb)	3107.54	3131.35
Effective sequences near target(Mb)	1201.50	1138.46
Effective sequences on or near target(Mb)	4309.04	4269.80
Number of reads uniquely mapped to target	34072767	34335223
Number of reads uniquely mapped to genome	62289651	45547432
Average sequencing depth on target	60.29	60.75
Average sequencing depth near target	16.40	15.54
Mismatch rate in target region	0.34%	0.33%
Mismatch rate in all effective sequence	0.34%	0.30%
Base covered on target	50743793	50599905
Coverage of target region	98.4%	98.2%
Base covered near target	69618007	67761541
Coverage of flanking region	95.0%	92.5%
Fraction of target covered with at least 20x	78.9%	79.4%
Fraction of target covered with at least 10x	88.6%	88.6%
Fraction of target covered with at least 4x	94.9%	94.6%
Fraction of flanking region covered with at least 20x	26.8%	25.5%
Fraction of flanking region covered with at least 10x	49.0%	46.0%
Fraction of flanking region covered with at least 4x	76.7%	71.9%
Mapping rate	98.66%	99.39%
Duplicate rate	3.60%	4.26%

Table 4.8 Family 2's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference

genome. The values of the two individuals are very close, indicating that the read alignment process was consistent.

Variant calling statistics

After SNP and indels are identified (as described in section 2.4.2), ANNOVAR and VEP was used to annotate and classify each variant. The table below display summary statistics for SNPs. 'Hom' means homozygous and 'Het' means heterozygous. The term 'Exonic' refers only to the coding exonic portion (also referred to as 'CDS' in Figure 4.6 and other subsequent related figures); thus does not include UTRs, which are named separately as UTR5 and UTR3). 'Splicing' in ANNOVAR is defined as a variant that is within 2 bp away from an exon/intron boundary. The terms "Upstream" and "Downstream" are defined as 1 kb away from the transcription start site and end site respectively. 'Ti' refers to Transitions* and 'Tv' refers to Transversions†. The length distribution of the InDels in whole target region and CDS region are also plotted below. Details about the variant calling process can be found at section 2.4.2.

Two complementary indel distribution figures are presented for each individual. The first figure is for indels falling within the coding regions (which is where the frameshifting in the amino acid sequence occurs); and the other is for all the indels (including non-coding regions).

The explanation above will only be stated once and will apply for all tables and figures representing variant calling statistics in all families.

Individual 1

Tables 4.9 and 4.10 summarises the SNV and indels detected and called for individual 1 respectively. Figures 4.4 and 4.5 depict the indel and CNV distribution throughout the genome as captured by WES platform.

* Point mutations which change a purine (i.e. adenine or guanine) nucleotide to another purine

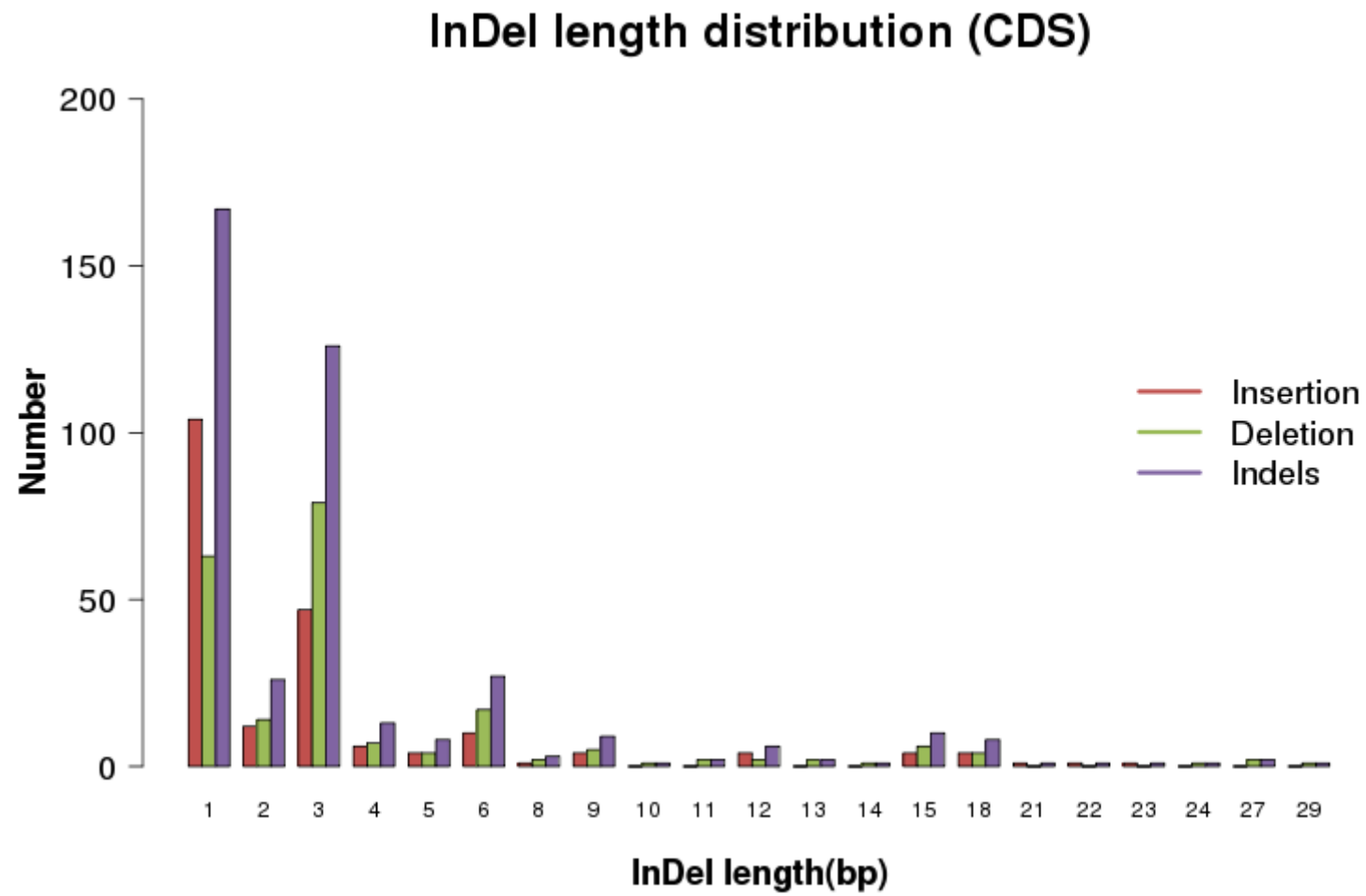
† Mutation of a (two ring) purine to a (one ring) pyrimidine (i.e. cytosine or thymine) or vice versa

Categories	Value	Categories	Value
Total	80084	Splicing	55
1000genome and dbsnp135	75933	NcRNA	2749
1000genome specific	192	UTR5	910
dbSNP135 specific	2432	UTR5 and UTR3	2
dbSNP rate	97.85%	UTR3	2598
Novel	1527	Intronic	49886
Hom	38296	Upstream	637
Het	41788	Upstream and downstream	65
Synonymous	9651	Downstream	454
Missense	8370	Intergenic	4609
Stopgain	61	SIFT	1004
Stoploss	37	Ti/Tv	2.4440
Exonic	17867	dbSNP Ti/Tv	2.4510
Exonic and splicing	252	Novel Ti/Tv	2.2420

Table 4.9 Family 1 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9227	Stopgain	4
1000genome and dbsnp135	4655	Stoploss	0
1000genome specific	943	Exonic	388
dbSNP135 specific	1819	Exonic and splicing	3
dbSNP rate	70.16%	Splicing	30
Novel	1810	NcRNA	320
Hom	4655	UTR5	75
Het	4572	UTR5 and UTR3	0
Frameshift Insertion	104	UTR3	392
Non-frameshift Insertion	72	Intronic	7411
Frameshift Deletion	98	Upstream	81
Non-frameshift Deletion	113	Upstream and downstream	7
Frameshift block substitution	0	Downstream	53
Non-frameshift block substitution	0	Intergenic	467

Table 4.10 Family 1 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.9 above for a comparison.



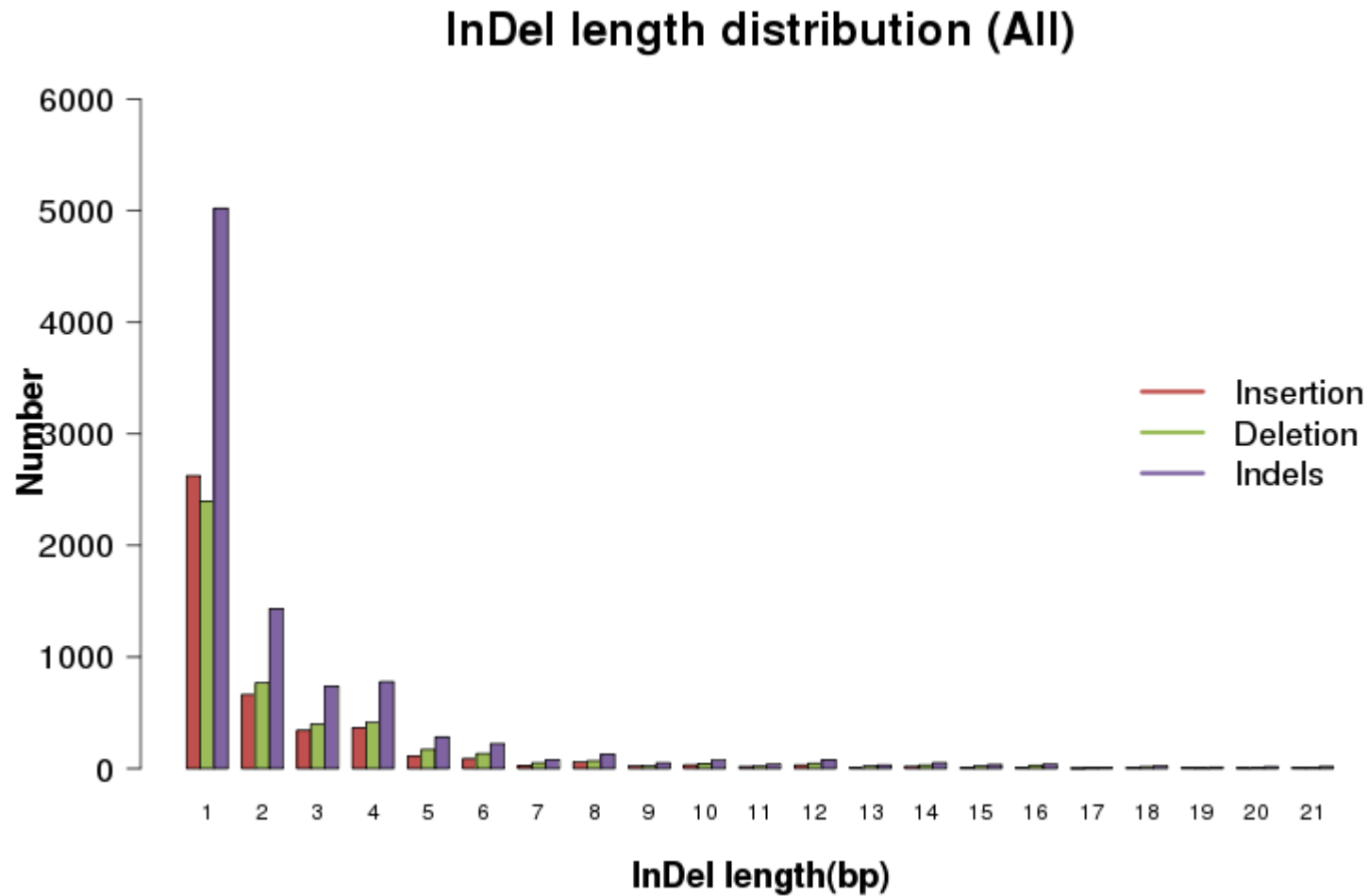


Figure 4.4 Family 1 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

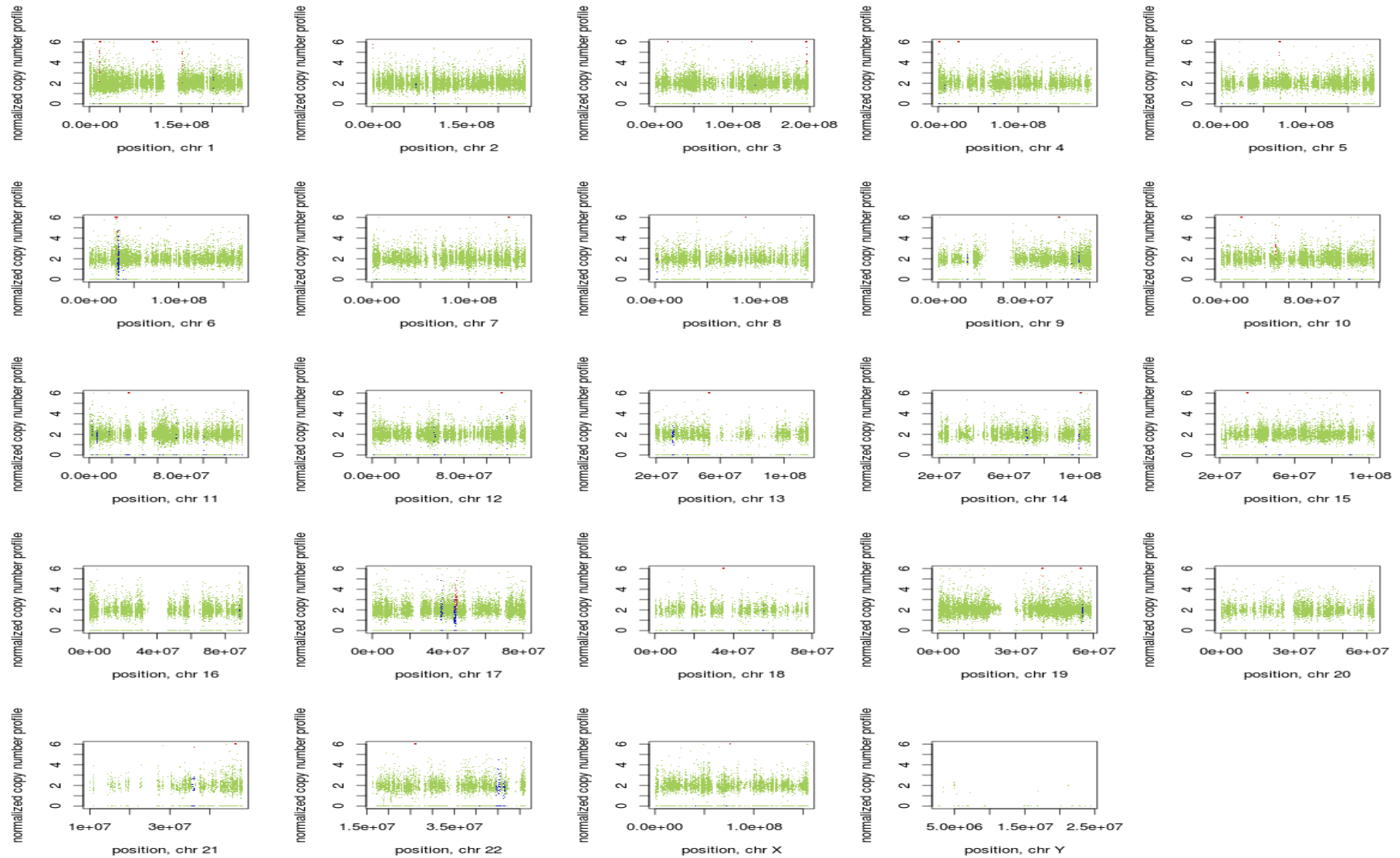


Figure 4.5 Family 1 Individual 1 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels. None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality issues.

Individual 2

Tables 4.11 and 4.12 summarises the SNV and indels detected and called for individual 2 respectively. Figures 4.6 and 4.7 depict the indel and CNV distribution throughout the genome as captured by WES platform.

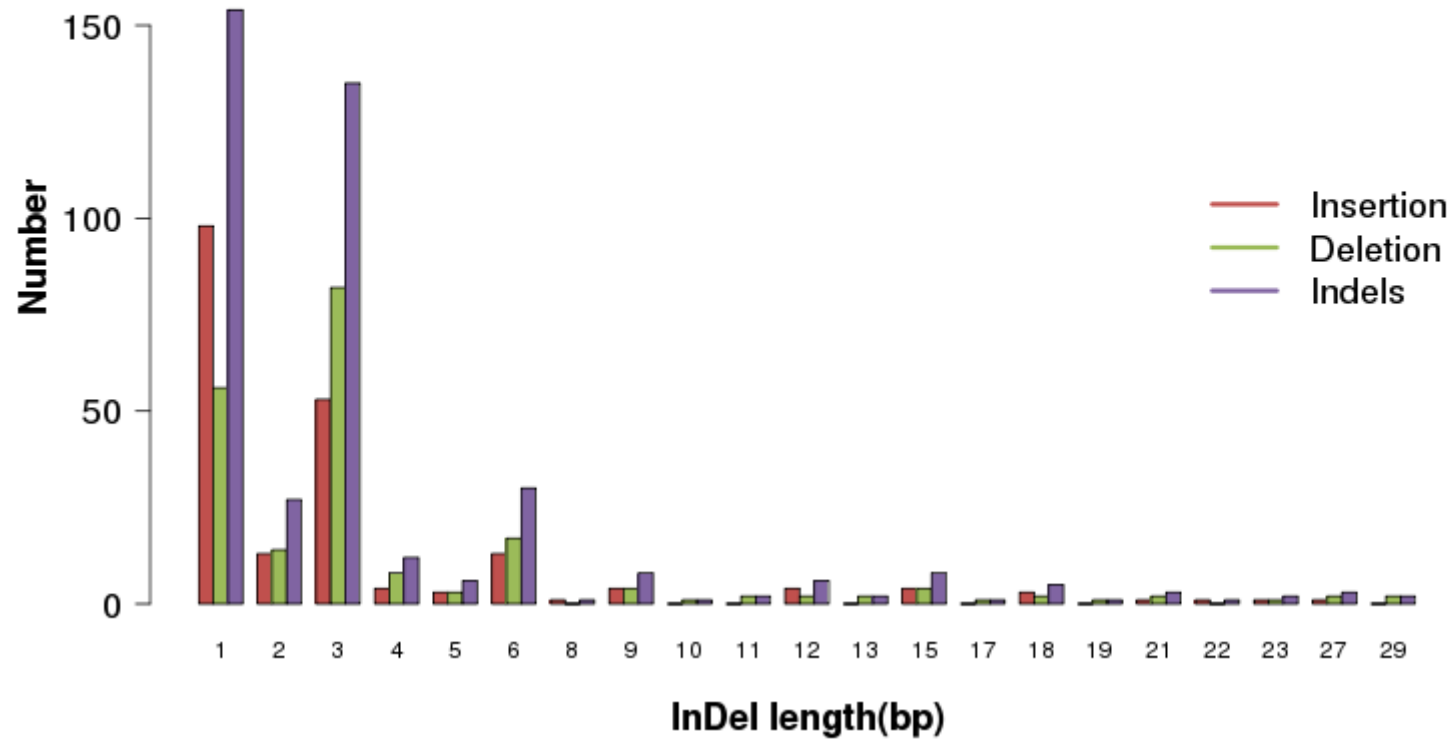
Categories	Value	Categories	Value
Total	76734	Splicing	58
1000genome and dbsnp135	72670	NcRNA	2594
1000genome specific	210	UTR5	872
dbSNP135 specific	2351	UTR5 and UTR3	2
dbSNP rate	97.77%	UTR3	2485
Novel	1503	Intronic	47247
Hom	36808	Upstream	542
Het	39926	Upstream and downstream	54
Synonymous	9445	Downstream	420
Missense	8343	Intergenic	4576
Stopgain	63	SIFT	944
Stoploss	33	Ti/Tv	2.4425
Exonic	17632	dbSNP Ti/Tv	2.4505
Exonic and splicing	252	Novel Ti/Tv	2.1444

Table 4.11 Family 1 Individual 2's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	8780	Stopgain	3
1000genome and dbsnp135	4393	Stoploss	0
1000genome specific	923	Exonic	372
dbSNP135 specific	1749	Exonic and splicing	3
dbSNP rate	69.95%	Splicing	38
Novel	1715	NcRNA	283
Hom	4410	UTR5	78
Het	4370	UTR5 and UTR3	0
Frameshift Insertion	92	UTR3	359
Non-frameshift Insertion	81	Intronic	7060
Frameshift Deletion	88	Upstream	65
Non-frameshift Deletion	111	Upstream and downstream	5
Frameshift block substitution	0	Downstream	59
Non-frameshift block substitution	0	Intergenic	458

Table 4.12 Family 1 Individual 2's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.11 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

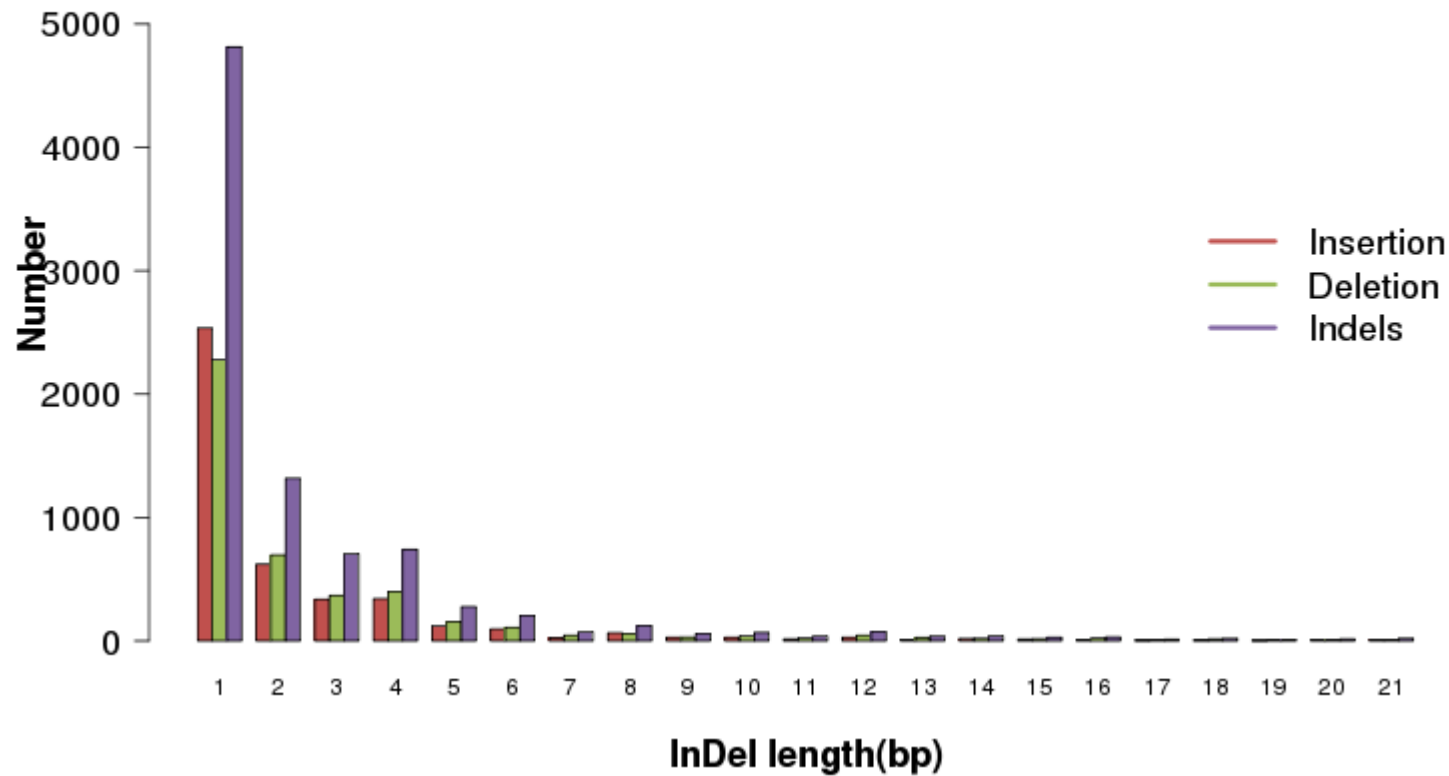


Figure 4.6 Family 1 Individual 2's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

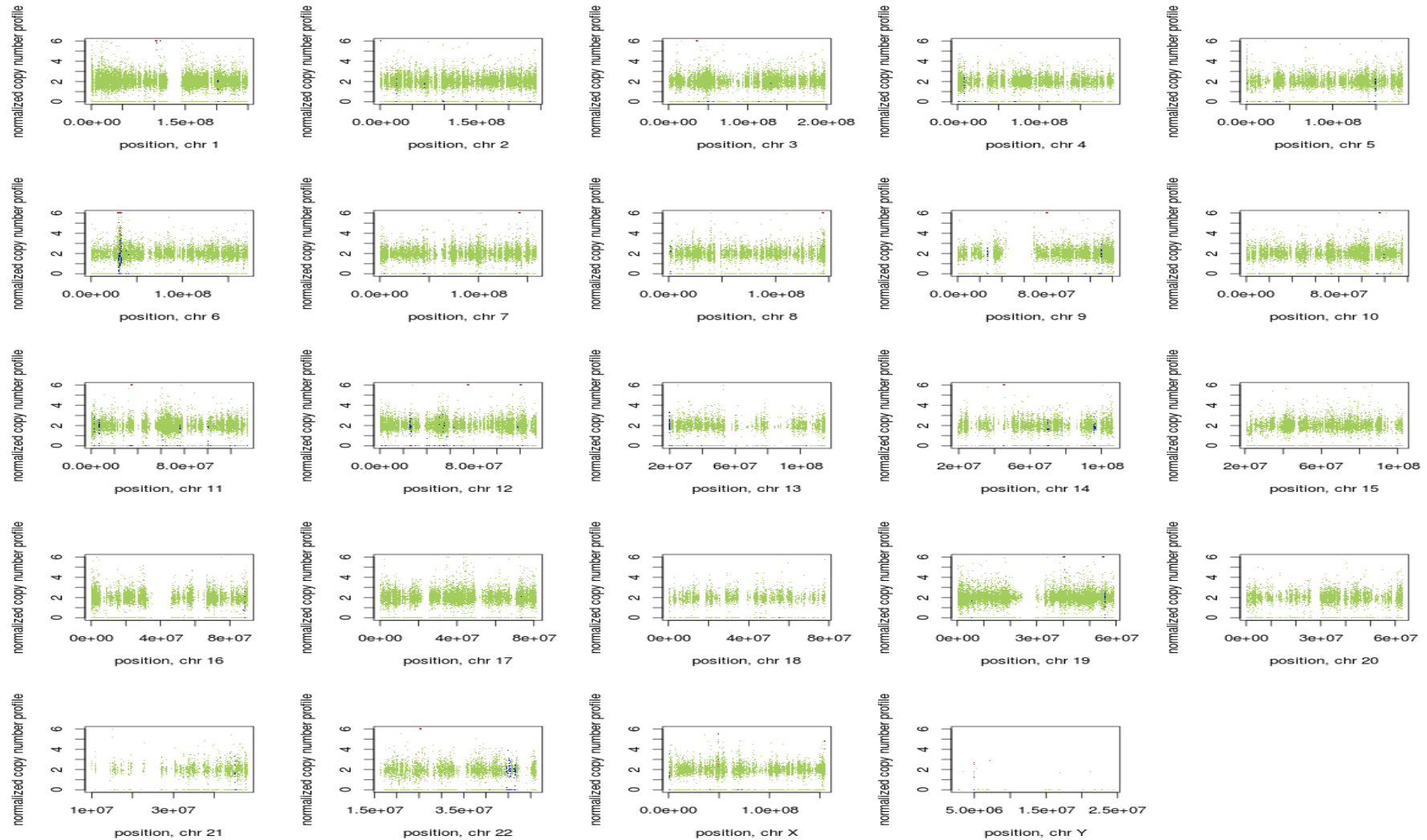


Figure 4.7 Family 1 Individual 2's CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels. None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality issues.

Analysis of whole-exome, LRoH and candidate genes

The two candidate gene lists and the SO terms list were used to filter and prioritise genes and variants with these genes (details in section 4.4.7).

Initially, all LRoH regions identified by Plink [294] and our custom Python script were reviewed in IGV [284] to check whether any known PCD gene (or reported linkage region, see section 4.1.2 for details) resides within the autozygous regions (Figures 4.8 and 4.9). None of these regions spanned a known human PCD gene or a reported linkage region. Several filters (e.g. MAF, consequence of variant) were applied systematically on all mutations to single out any potentially causal ones. All remaining Φ mutations which were in a homozygous state were analysed separately in the two candidate lists. This yielded 13 homozygous Φ mutations in list 1 (i.e. known genes) and a total of 370 Φ variants in list 2 (i.e. all suspected ciliome genes) in individual 1. However all the mutations in list 1 were common (i.e. with MAF = >1%) and only 4 of the mutations in list 2 passed this MAF criteria, which were a stop gained (p.R263*, see Figure 4.15) in the *DNALI1* gene (Ensembl ID: ENSG00000163879) and the other three were missense mutations in the *NUP155*, *RFX1* and *ZFHX4* genes. The stop gain (i.e. p.R263*) and the missense mutation in *NUP155* were absent in dbSNP, EVS, our internal database and in the 1000GP databases (the stop gain is present in 2 heterozygotes out of 107,844 in the ExAC database); but the other remaining missense mutations were present and relatively common (>0.1%) in EVS and our internal database. However none of these mutations were homozygous in the other affect sibling. The filtering process is depicted in Figure 4.16.

The stop gain was located within a long LRoH region (~10Mb) on chromosome 1 of individual 1. The nonsense mutation falls within exon 5 of *DNALI1* which has 6 exons in total and codes for 280 residues where it is expected to deem the gene product (i.e. protein) dysfunctional. The variant resides in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 570.9 (also see Table 4.13) [310]. STRING was used to predict the *DNALI1* interactome which yielded

many already identified PCD causal genes (Figure 4.12) [258]. A final analysis was carried out on all the mutations outside of the two candidate lists and no additional mutations were identified which met all the criteria set out in Figure 4.16. The CNV analysis also yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

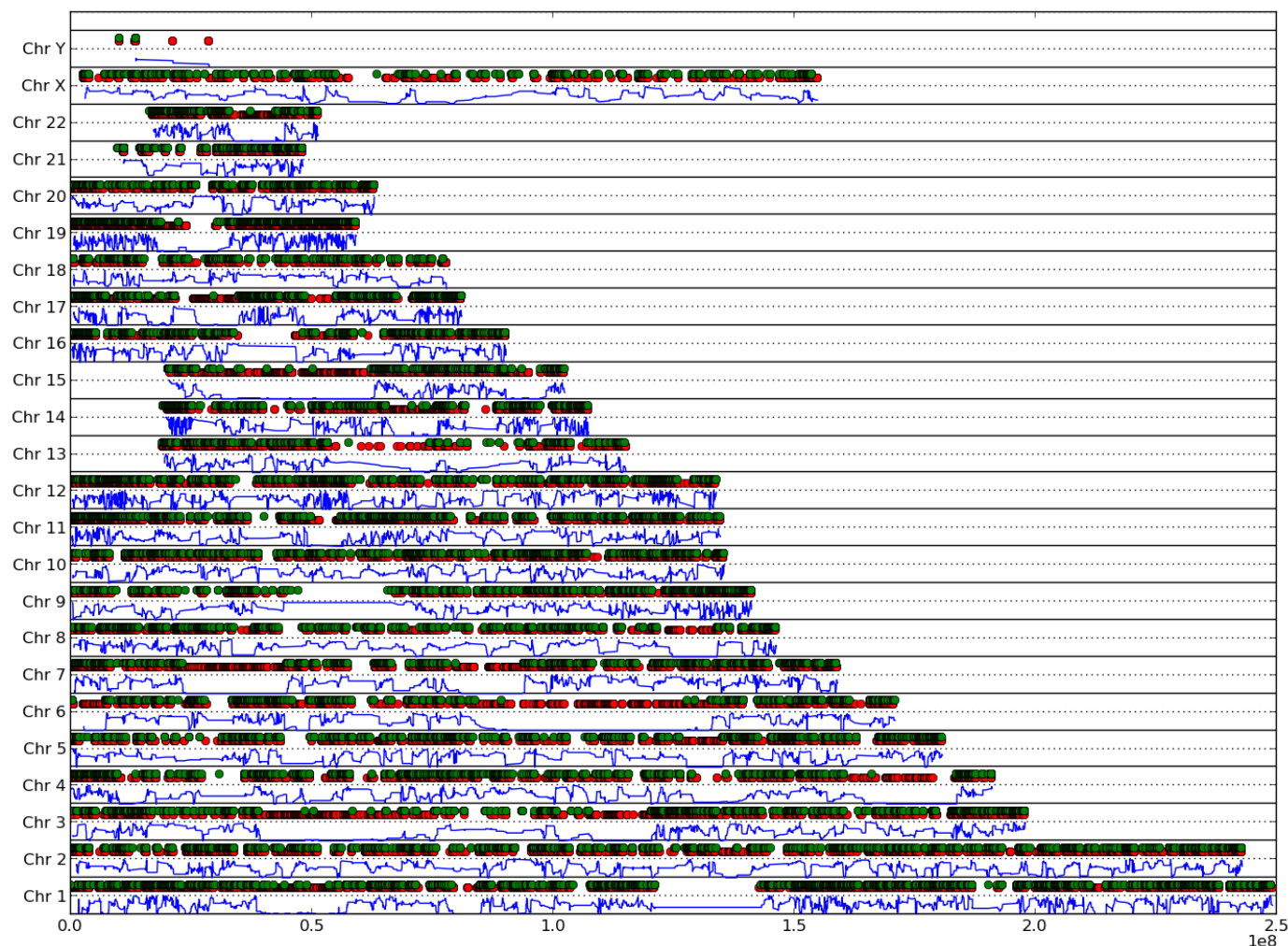


Figure 4.8 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 7 has an autozygous region ~20Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter. NB: These images may have been edited to ensure confidentiality/anonymity of the participants. Some LRoHs may also have been shortened/extended for the same reason.

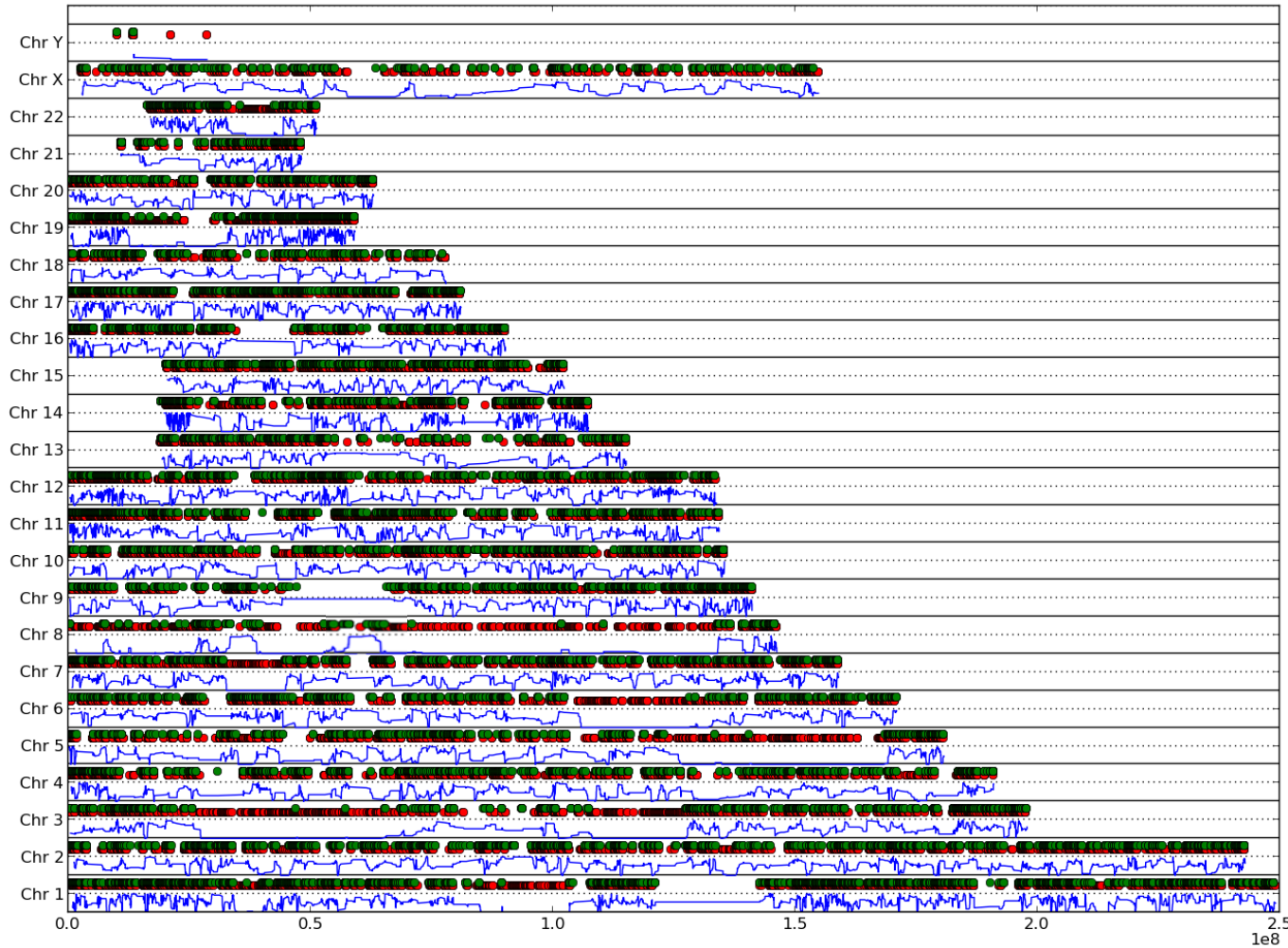


Figure 4.9 Plotting of variant status across the genome for individual 2. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 8 has an autozygous region ~50Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

Modelling mutation effect on protein structure

Uploading the amino acid sequences stated in section 4.4.9 to the Robetta server yielded 5 different models for both the mutant and the wild type DNALI1 protein. Figures 4.10 and 4.11 below depict the first model for both. The affected residue is labelled in both.

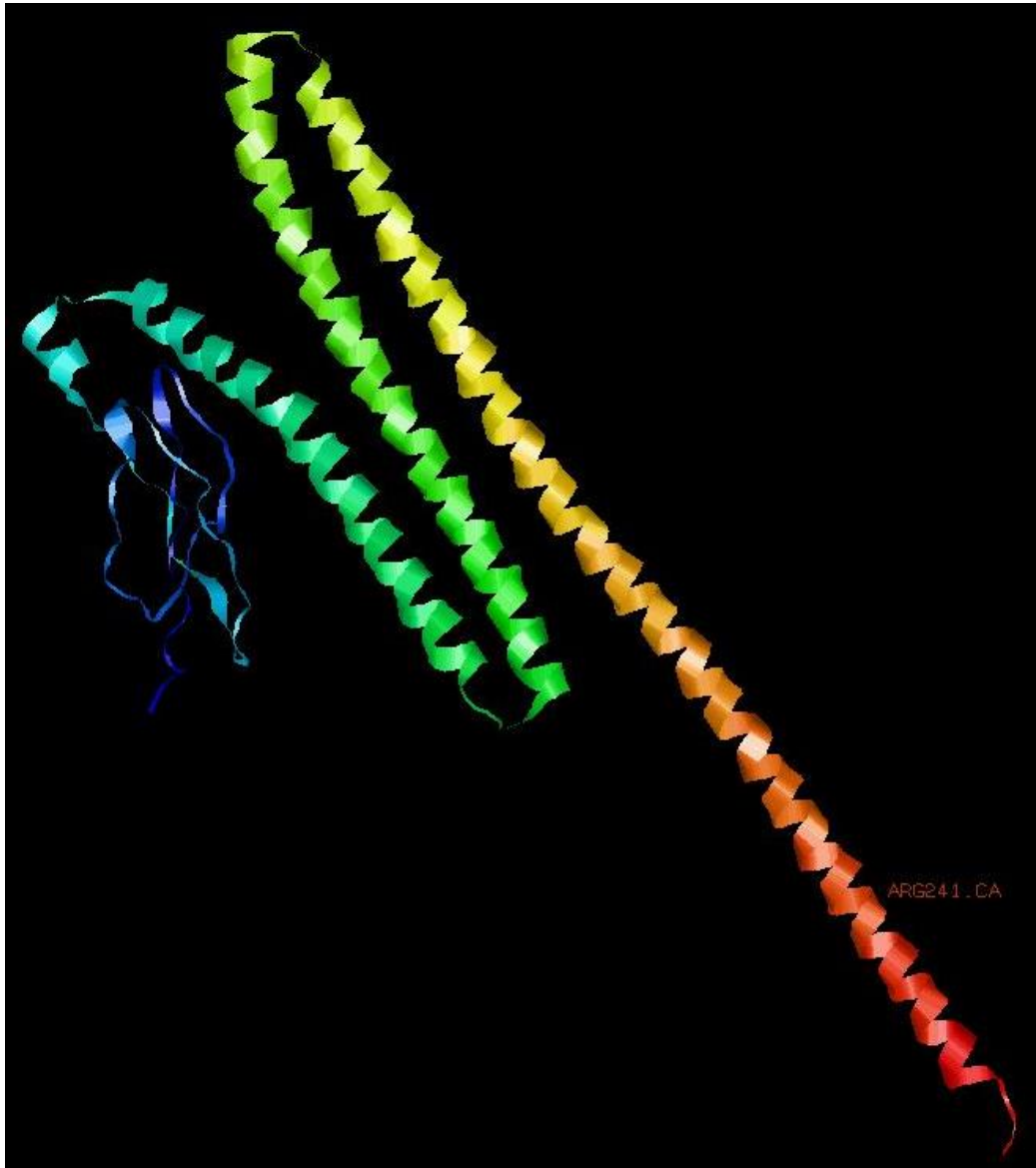


Figure 4.10 Protein structure of wild type DNALI1 protein. Only the first (most likely) model is shown and the location of the mutation (p.R263*) is labelled. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.



Figure 4.11 Protein structure of mutant DNALI1 protein. Only the first model is shown. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol. The mutation p.R263* cannot be presented as the protein product is truncated from the stopgain locus (see Figure 4.10). Comparing the predicted protein structures of the wild type and mutant forms of the DNALI1 protein shows that most of the protein remains the same in the latter. See discussion on section 4.6.2 for details.

Ultrastructure of respiratory cilia

The EM image of the cross-sections of the mutant and normal respiratory cilia are depicted in Figures 4.14A-E. The EM images clearly show the lack of subfiber B, IDA and ODA (although 9+2 formation is retained) in the mutant cilia in relation to the ones from a control.

Start	Amino acid sequence	End	Entry	Entry Name	Organism
240	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	280	O14645	IDLC_HUMAN	Homo sapiens (Human)
240	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	280	G3QXI9	G3QXI9_GORGO	Gorilla gorilla gorilla (Lowland gorilla)
240	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	280	Q0IIN3	Q0IIN3_MOUSE	Mus musculus (Mouse)
240	EKREAERRQVEEKKHAE EIQFLK R TNQQLKAQLEGIIAPKK	280	Q28IW2	Q28IW2_XENTR	Xenopus tropicalis (Western clawed frog)
223	EDEERARREEEERKHTEEVAFRR R TYETLPRICRLS-----	263	S9WQU2	S9WQU2_9TRYP	Angomonas deanei
241	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	281	M3WA65	M3WA65_FELCA	Felis catus (Cat)
239	EKRETEERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	279	F7DKP4	F7DKP4_HORSE	Equus caballus (Horse)
242	EKRESEKRQVEEKKRHNEEIQFLK R TNQQLKAQLEGIIAPKK	282	G1T058	G1T058_RABIT	Oryctolagus cuniculus (Rabbit)
242	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	282	I3LJ86	I3LJ86_PIG	Sus scrofa (Pig)
262	EKRESERRQVEEKKHNEEIQFLK R TNQQLKAQLEGIIAPKK	302	H2PYP2	H2PYP2_PANTR	Pan troglodytes (Chimpanzee)

Table 4.13 Local sequence alignment containing the mutated residue from multiple alignment of the *DNALI1* gene in different organisms (relevant species shown). The highlighted Arginine (R) residue is found to be highly conserved across many species where the gene is predicted to be a homologue of the human *DNALI1* gene. The alignment was carried out using the Uniprot website's BLAST and Align functions (<http://www.uniprot.org>) [413].

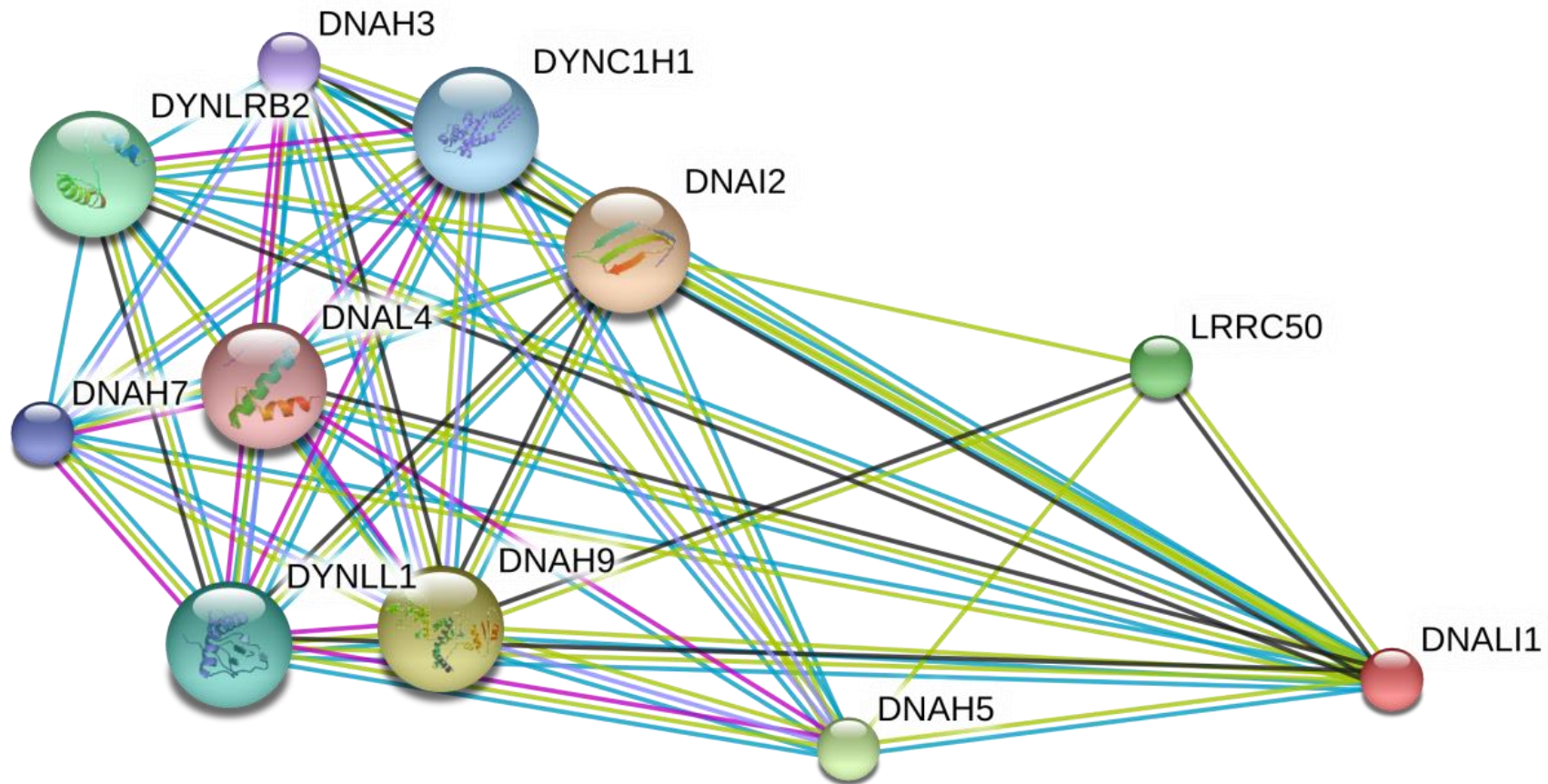


Figure 4.12 Proteins DNALI1p is predicted to interact with (top 10 shown). *DNAH5*, *DNAI2* and *LRRC50* (also known as *DNAAF1*) have previously been identified as human PCD genes. The predictions were made and the image was generated using STRING (v9.1) [258].

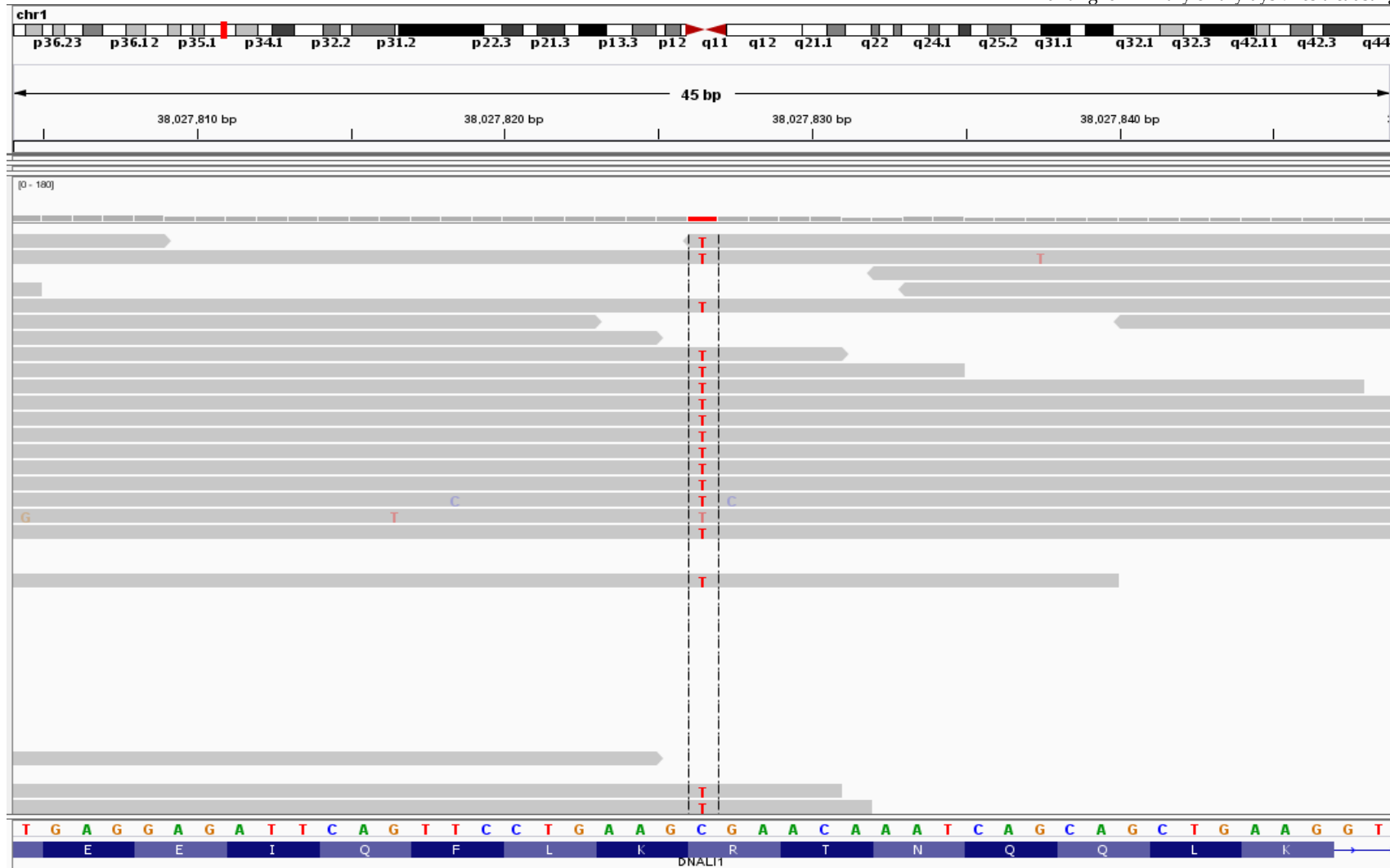
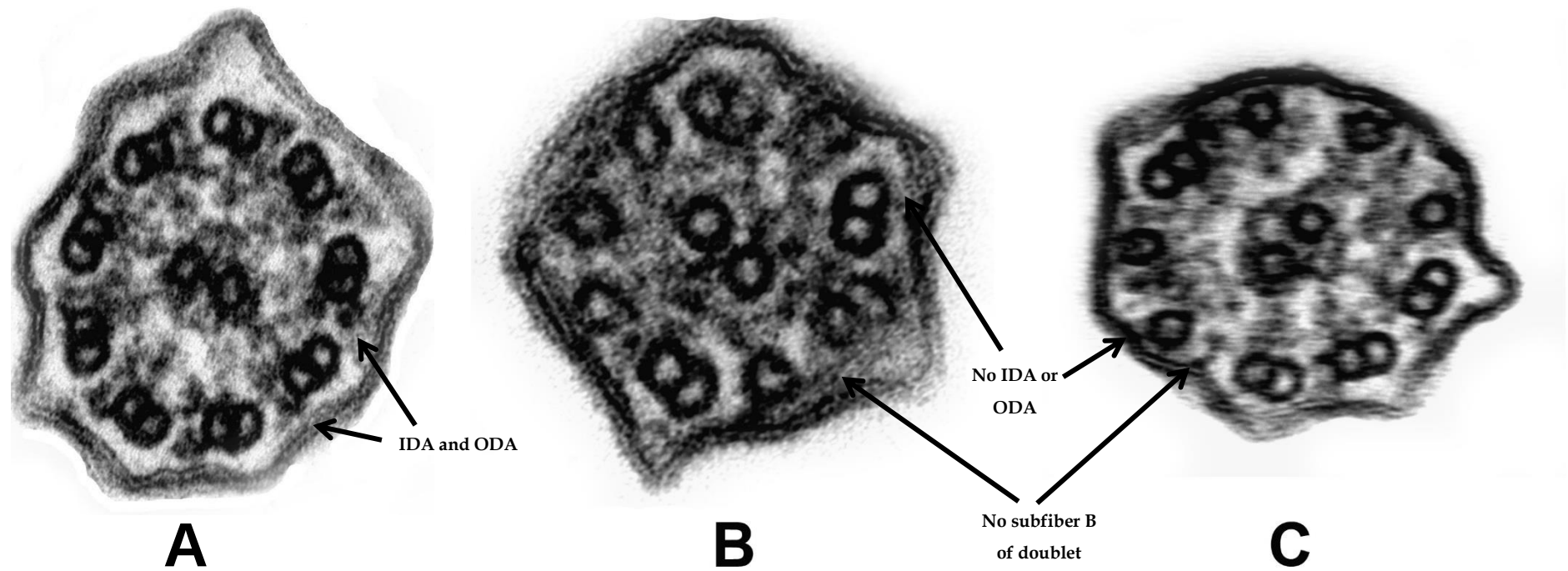
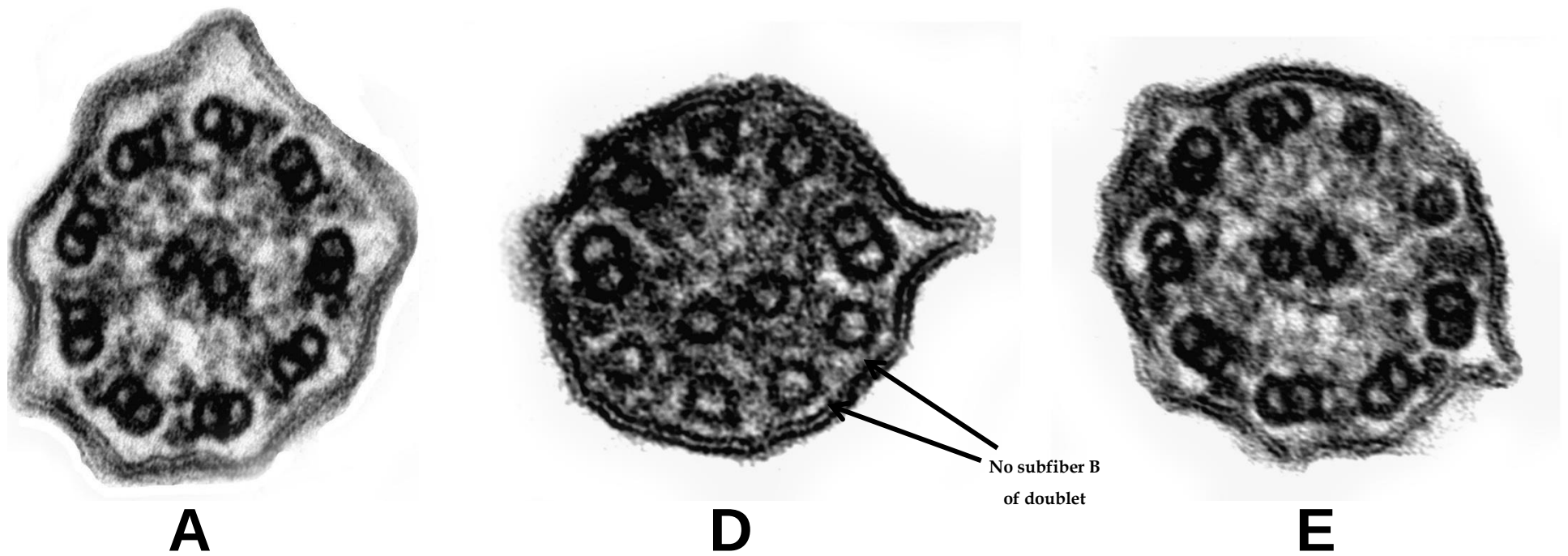


Figure 4.13 Reads mapped to the reference human genome hg19 at the site of the p.R263* mutation. All nineteen reads are high quality (18 reads shown, one further below) and there are no wild type alleles. The image was created using IGV [284].



Figures 4.14A-C EM image of (A) Control cilia (B and C) individual 1's cilia. Note the lack of subfiber B, IDA and ODA in the mutated cilia, although 9+2 formation is retained. EM magnification: 250000x



Figures 4.14D-E EM image of (A) Control cilia (D and E) individual 2's cilia. Note the lack of subfiber B, IDA and ODA in the mutated cilia, although 9+2 formation is retained. EM magnification: 250000x

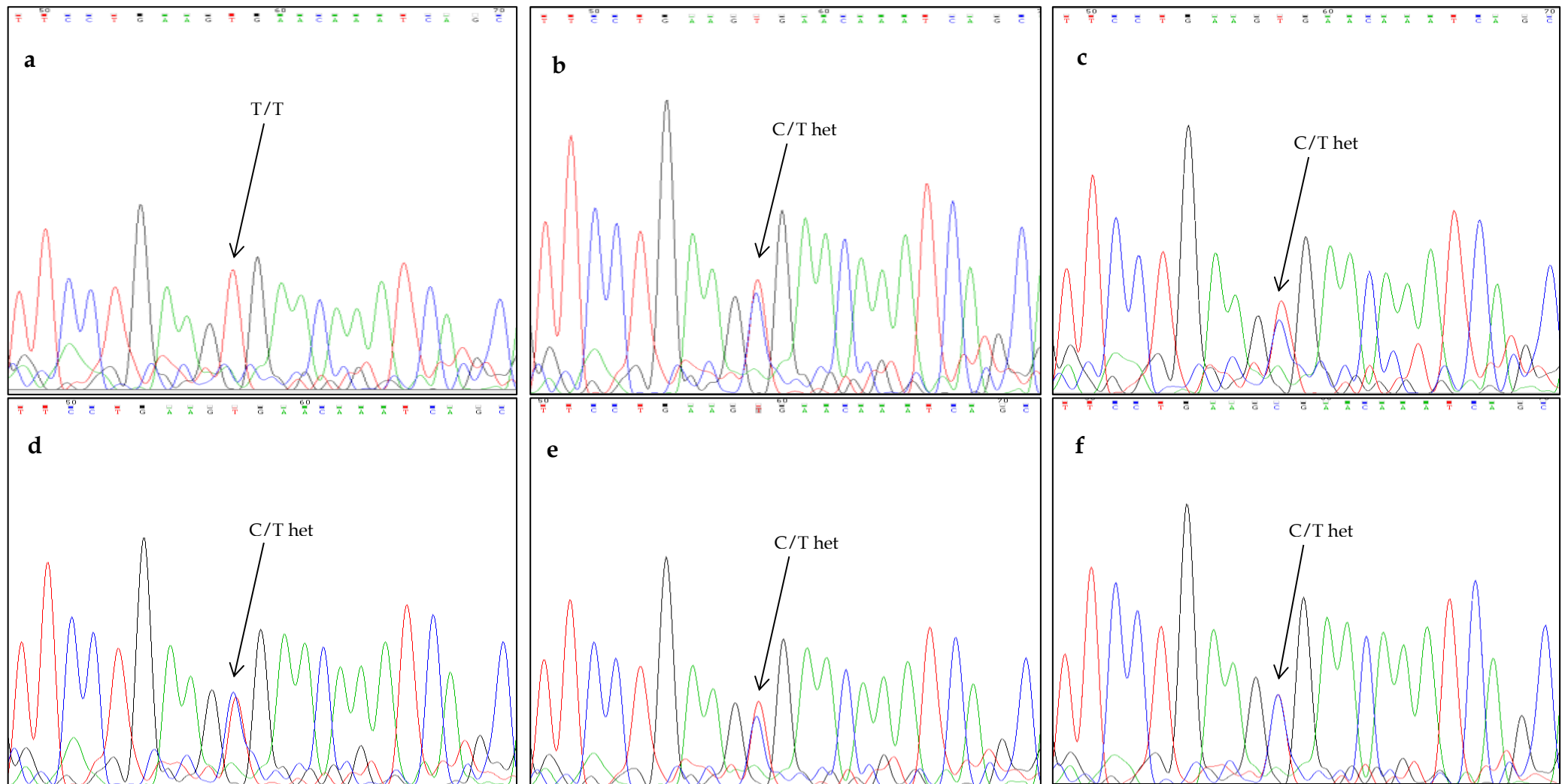


Figure 4.15 Confirmation of variant status in proband and other family members using *Sanger sequencing. (a) Proband (b) Affected sister (c) Unaffected brother (d) Mother (e) Father (f) Unaffected brother. DNA Chromatogram images were created using Chromas Lite (v2.1.1) . *Peak height imbalances could have been caused by low template DNA, degraded DNA and/or preferential amplification. The affected sister is heterozygous, which is not compatible with autosomal recessive mode of inheritance.

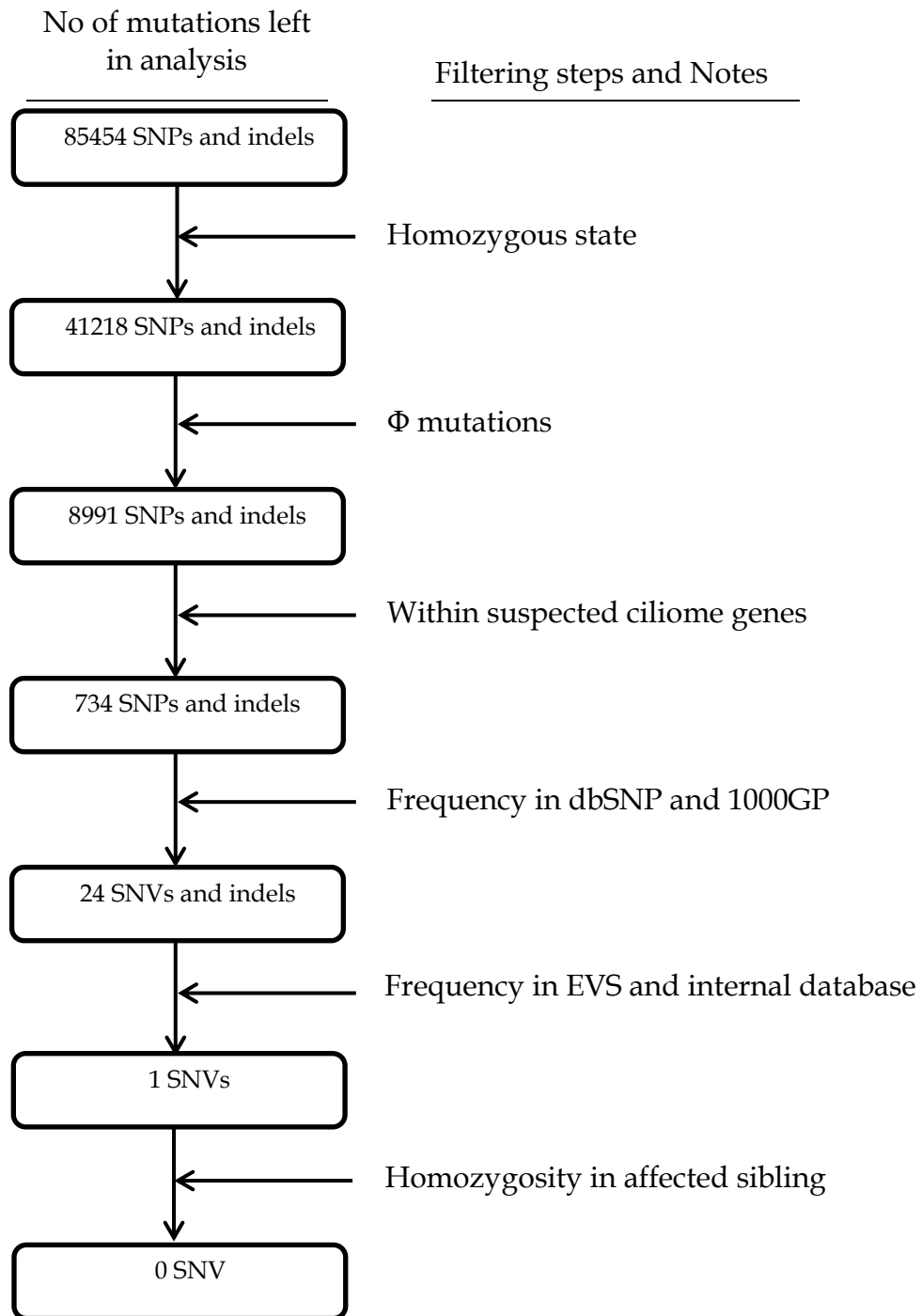


Figure 4.16 Filtering steps applied to all mutations in the exome of proband. After all the filtering steps in the above figure were applied, the total no of Φ variants was reduced to none.

4.5.3. Family 2

This section will present the results for family 2 and describe the findings:

No of whole-exome sequenced participants: 3

No of other family members with DNA available: 2 (i.e. both parents)

WES data statistics

For individual 1, the total captured region was 117,665,039 base pairs (50,714,949 bases on target and 66,950,090 bases near target, the latter being flanking regions within 200bp of exons). Coverage of target (i.e. exons) and flanking regions (e.g. introns, splice sites) was 98.4% and 91.4% respectively. The average sequencing depth on target was 61.08 and the fraction of target covered with at least 20 and 10 reads was 79.4% and 88.7% respectively (and >4 read depth = 94.8%). There were a total of 50,805,417 (high quality) reads with a mapping rate of 99.1%.

For individual 2, the total captured region was 119,076,103 base pairs (50,753,931 bases on target and 68,322,172 bases near target). Coverage of target and flanking regions was 98.5% and 93.3% respectively. The average sequencing depth on target was 61.04 and the fraction of target covered with at least 20 and 10 reads was 79.2% and 88.7% respectively (and >4 read depth = 94.8%). There were a total of 52,135,563 reads with a mapping rate of 99.35%.

For individual 3, the total captured region was 117,857,613 base pairs (50,648,396 bases on target and 67,209,217 bases near target). Coverage of target and flanking regions was 98.3% and 91.8% respectively. The average sequencing depth on target was 61.36 and the fraction of target covered with at least 20 and 10 reads was 80.5% and 89.37% respectively (and >4 read depth = 94.9%). There were a total of 50,753,856 reads with a mapping rate of 99.12%.

Table 4.14 summarises the data produced by the WES platform for the family.

Type	Individual 1	Individual 2	Individual 3
Number of Reads	54622530	53244014	53581070
Data Size	5462253000	5324401400	5358107000
N of fq1	28258	30050	78665
N of fq2	22236	21456	83452
GC(%) of fq1	44.57~44.67	44.93~45.02	45.35~45.35
GC(%) of fq2	44.7~44.8	45.06~45.14	45.45~45.45
Q20(%) of fq1	96.42~96.62	96.38~96.62	97.49~97.76
Q20(%) of fq2	94.91~95.06	95.14~95.28	96.58~96.82
Q30(%) of fq1	89.19~89.68	89.16~89.69	92.01~92.71
Q30(%) of fq2	87.01~87.40	87.39~87.74	90.79~91.28

Table 4.14 Family 2's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values and close to each other between the three individuals.

Read alignment statistics

Table 4.15 (next page) summarises the quality of alignment of the reads produced by WES platform to the reference genome for family 2.

Exome Capture Statistics	Individual 1	Individual 2	Individual 3
Initial bases on target	51543125	51543125	51543125
Initial bases near target	73252334	73252334	73252334
Initial bases on or near target	124795459	124795459	124795459
Total effective reads	50805417	52135563	50753856
Total effective yield(Mb)	4992.23	5125.43	5006.95
Average read length(bp)	98.26	98.31	98.65
Effective sequences on target(Mb)	3148.50	3146.42	3162.90
Effective sequences near target(Mb)	1074.35	1174.79	1087.77
Effective sequences on or near target(Mb)	4222.85	4321.20	4250.67
Number of reads uniquely mapped to target	34078257	34321875	34219779
Number of reads uniquely mapped to genome	44804804	46163375	44942639
Average sequencing depth on target	61.08	61.04	61.36
Average sequencing depth near target	14.67	16.04	14.85
Mismatch rate in target region	0.34%	0.34%	0.24%
Mismatch rate in all effective sequence	0.32%	0.31%	0.23%
Base covered on target	50714949	50753931	50648396
Coverage of target region	98.4%	98.5%	98.3%
Base covered near target	66950090	68322172	67209217
Coverage of flanking region	91.4%	93.3%	91.8%
Fraction of target covered with at least 20x	79.4%	79.2%	80.5%
Fraction of target covered with at least 10x	88.7%	88.7%	89.3%
Fraction of target covered with at least 4x	94.8%	94.8%	94.9%
Fraction of flanking region covered with at least 20x	23.6%	26.2%	24.1%
Fraction of flanking region covered with at least 10x	42.7%	47.4%	43.8%
Fraction of flanking region covered with at least 4x	68.1%	73.4%	69.3%
Mapping rate	99.10%	99.35%	99.12%
Duplicate rate	3.71%	3.92%	4.43%

Table 4.15 Family 2's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference genome. The values of the three individuals are very close, indicating that the read alignment process was consistent.

Variant calling statistics

Individual 1

Tables 4.16 and 4.17 summarises the SNV and indels detected and called for individual 1. Figures 4.17 and 4.18 depict the indel and CNV distribution throughout the genome as captured by WES platform.

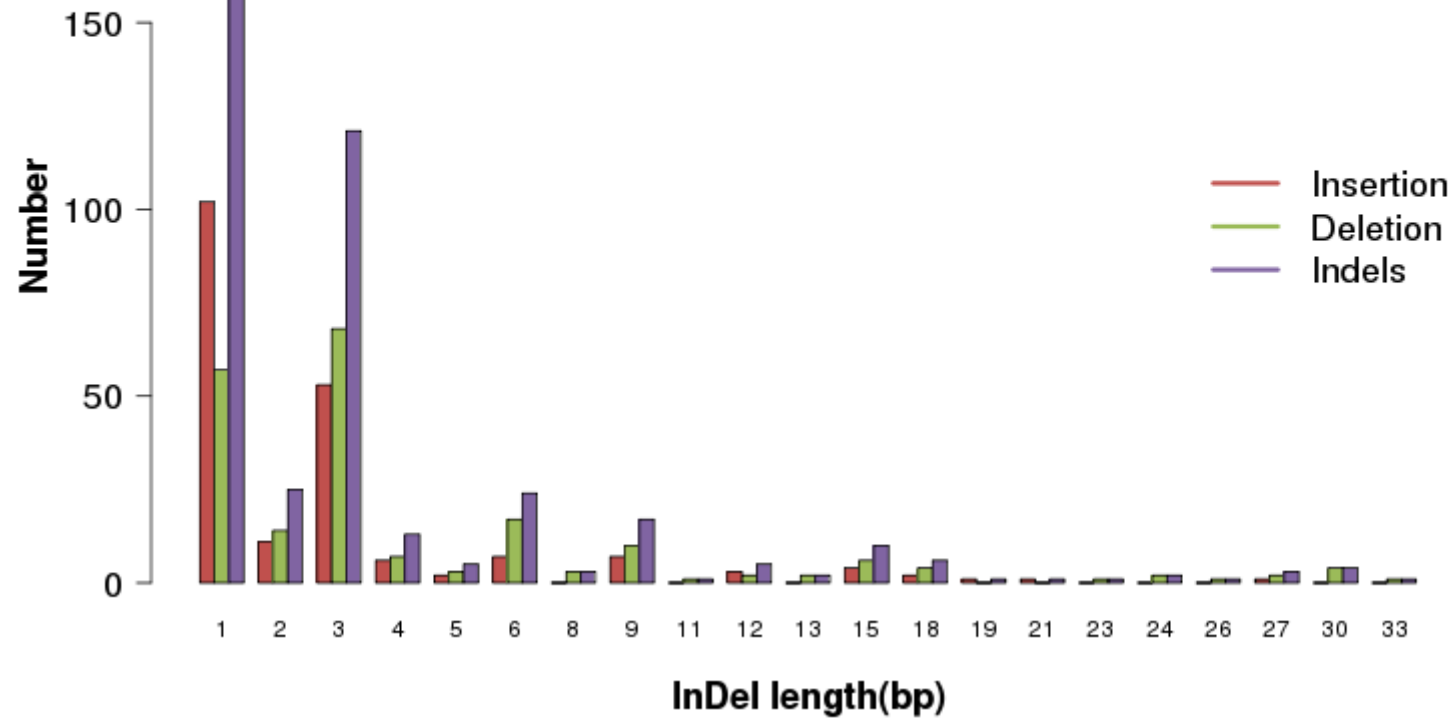
Categories	Value	Categories	Value
Total	75886	Splicing	57
1000genome and dbsnp135	71719	NcRNA	2538
1000genome specific	231	UTR5	841
dbSNP135 specific	2326	UTR5 and UTR3	2
dbSNP rate	97.57%	UTR3	2378
Novel	1610	Intronic	46350
Hom	36973	Upstream	559
Het	38913	Upstream and downstream	43
Synonymous	9620	Downstream	378
Missense	8449	Intergenic	4588
Stopgain	54	SIFT	1017
Stoploss	29	Ti/Tv	2.4140
Exonic	17913	dbSNP Ti/Tv	2.4160
Exonic and splicing	239	Novel Ti/Tv	2.3612

Table 4.16 Family 2 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	8621	Stopgain	3
1000genome and dbsnp135	4270	Stoploss	0
1000genome specific	944	Exonic	367
dbSNP135 specific	1690	Exonic and splicing	4
dbSNP rate	69.13%	Splicing	35
Novel	1717	NcRNA	298
Hom	4321	UTR5	79
Het	4300	UTR5 and UTR3	0
Frameshift Insertion	99	UTR3	372
Non-frameshift Insertion	76	Intronic	6890
Frameshift Deletion	81	Upstream	59
Non-frameshift Deletion	112	Upstream and downstream	5
Frameshift block substitution	0	Downstream	57
Non-frameshift block substitution	0	Intergenic	455

Table 4.17 Family 2 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.16 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

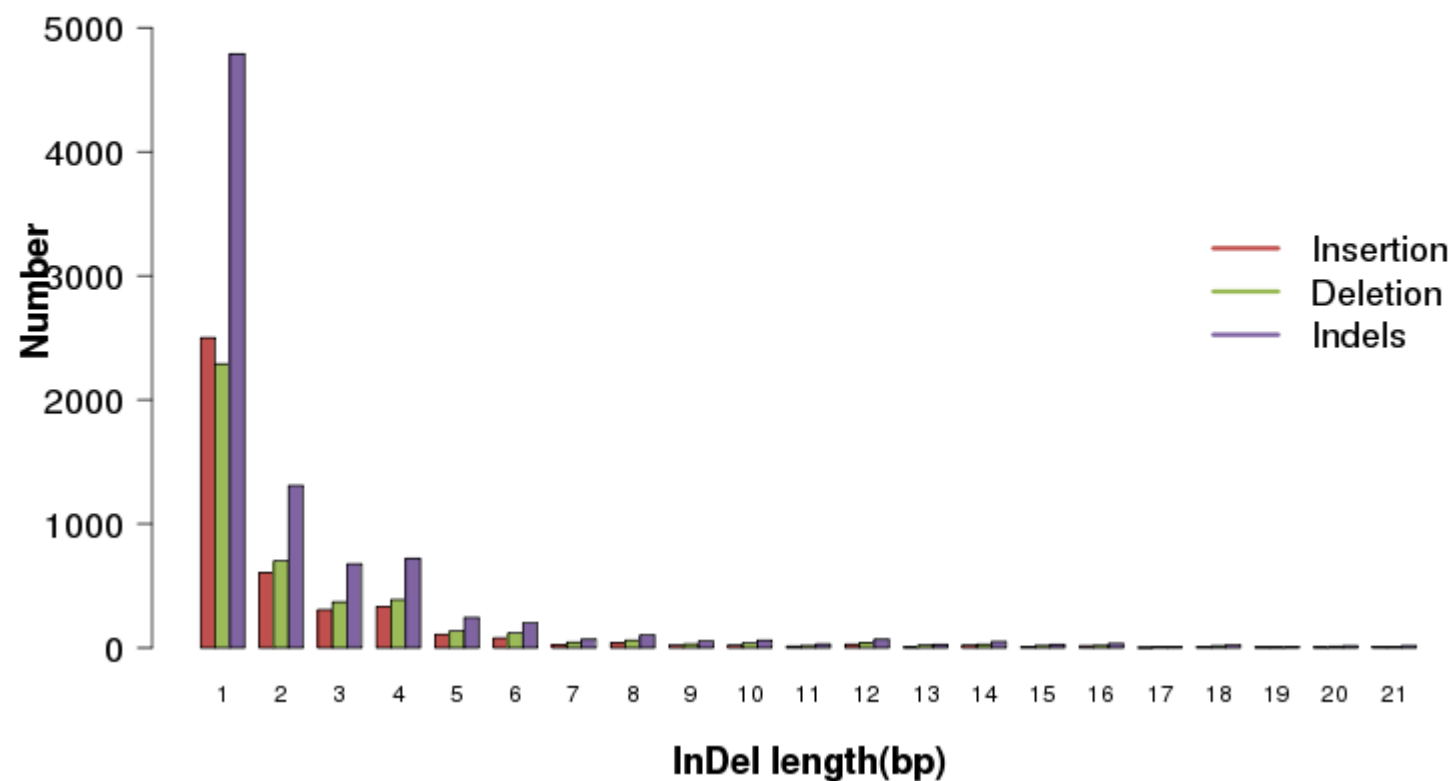


Figure 4.17 Family 2 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

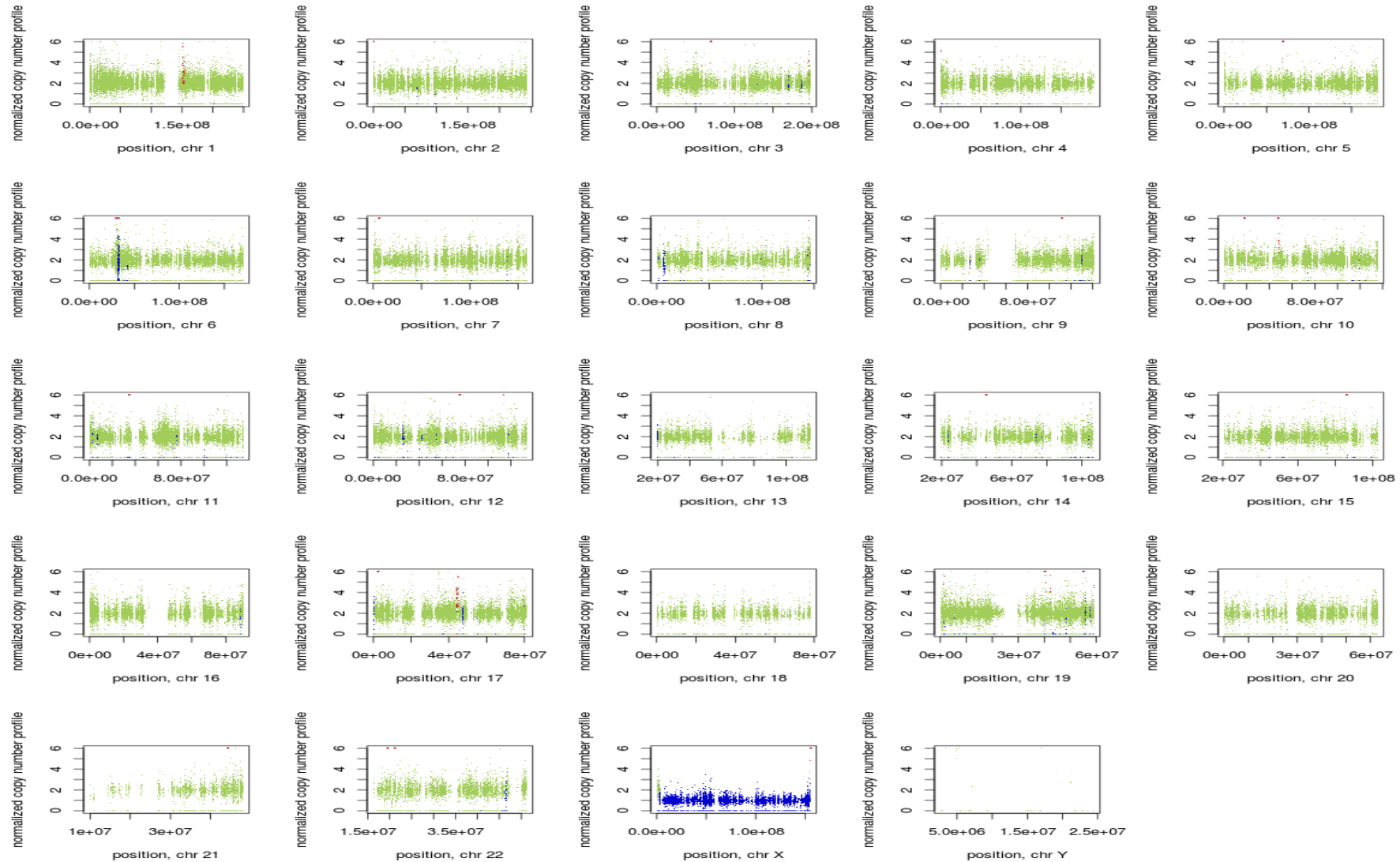


Figure 4.18 Family 2 Individual 1 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all blue chromosome X is expected when compared with a control female). None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Individual 2

Tables 4.18 and 4.19 summarises the SNV and indels detected and called for individual 2. Figures 4.19 and 4.20 depict the indel and CNV distribution throughout the genome as captured by WES platform.

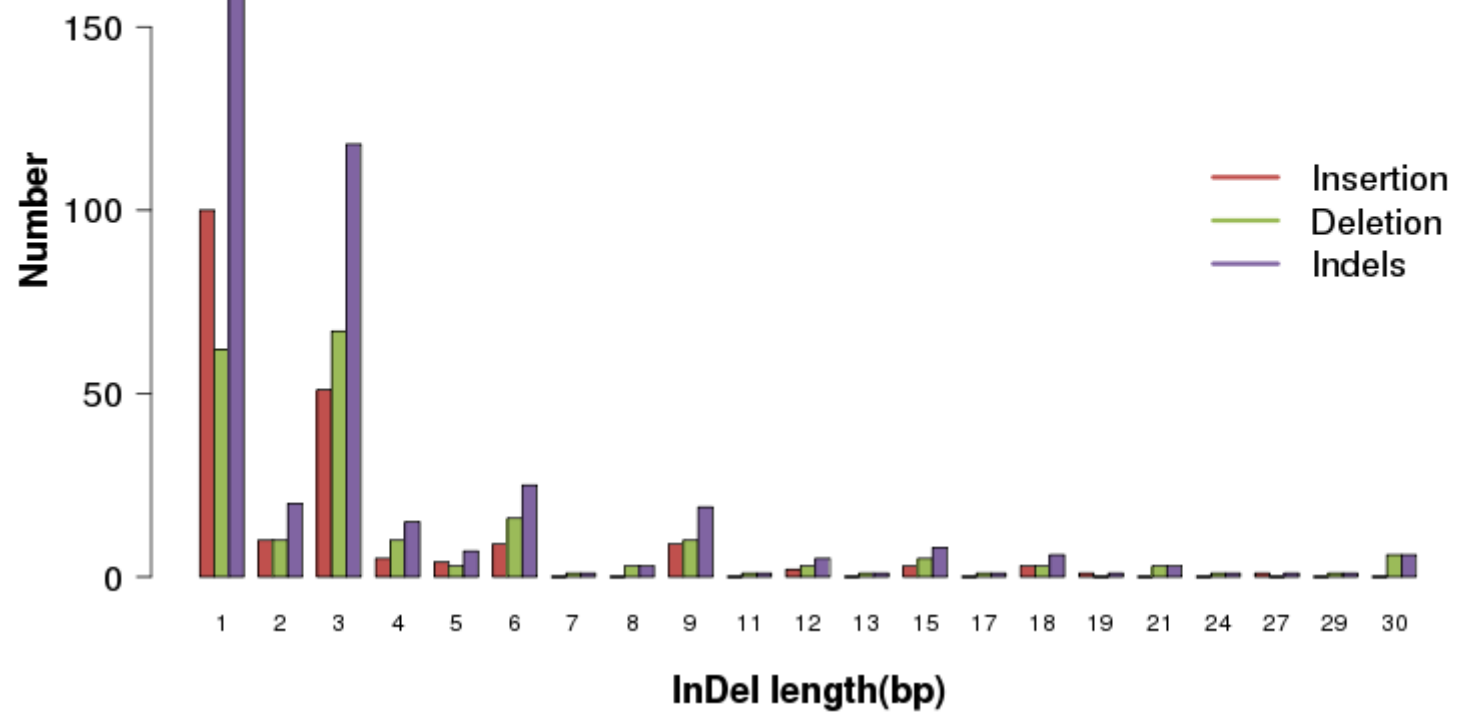
Categories	Value	Categories	Value
Total	80515	Splicing	57
1000genome and dbsnp135	76082	NcRNA	2724
1000genome specific	241	UTR5	905
dbSNP135 specific	2472	UTR5 and UTR3	3
dbSNP rate	97.56%	UTR3	2525
Novel	1720	Intronic	49944
Hom	36674	Upstream	603
Het	43841	Upstream and downstream	57
Synonymous	9795	Downstream	447
Missense	8613	Intergenic	4760
Stopgain	56	SIFT	1051
Stoploss	26	Ti/Tv	2.4109
Exonic	18243	dbSNP Ti/Tv	2.4206
Exonic and splicing	247	Novel Ti/Tv	2.1047

Table 4.18 Family 2 Individual 2's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9239	Stopgain	5
1000genome and dbsnp135	4599	Stoploss	0
1000genome specific	985	Exonic	365
dbSNP135 specific	1785	Exonic and splicing	5
dbSNP rate	69.10%	Splicing	36
Novel	1870	NcRNA	316
Hom	4421	UTR5	81
Het	4818	UTR5 and UTR3	0
Frameshift Insertion	93	UTR3	391
Non-frameshift Insertion	76	Intronic	7422
Frameshift Deletion	85	Upstream	71
Non-frameshift Deletion	111	Upstream and downstream	4
Frameshift block substitution	0	Downstream	67
Non-frameshift block substitution	0	Intergenic	481

Table 4.19 Family 2 Individual 2's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.18 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

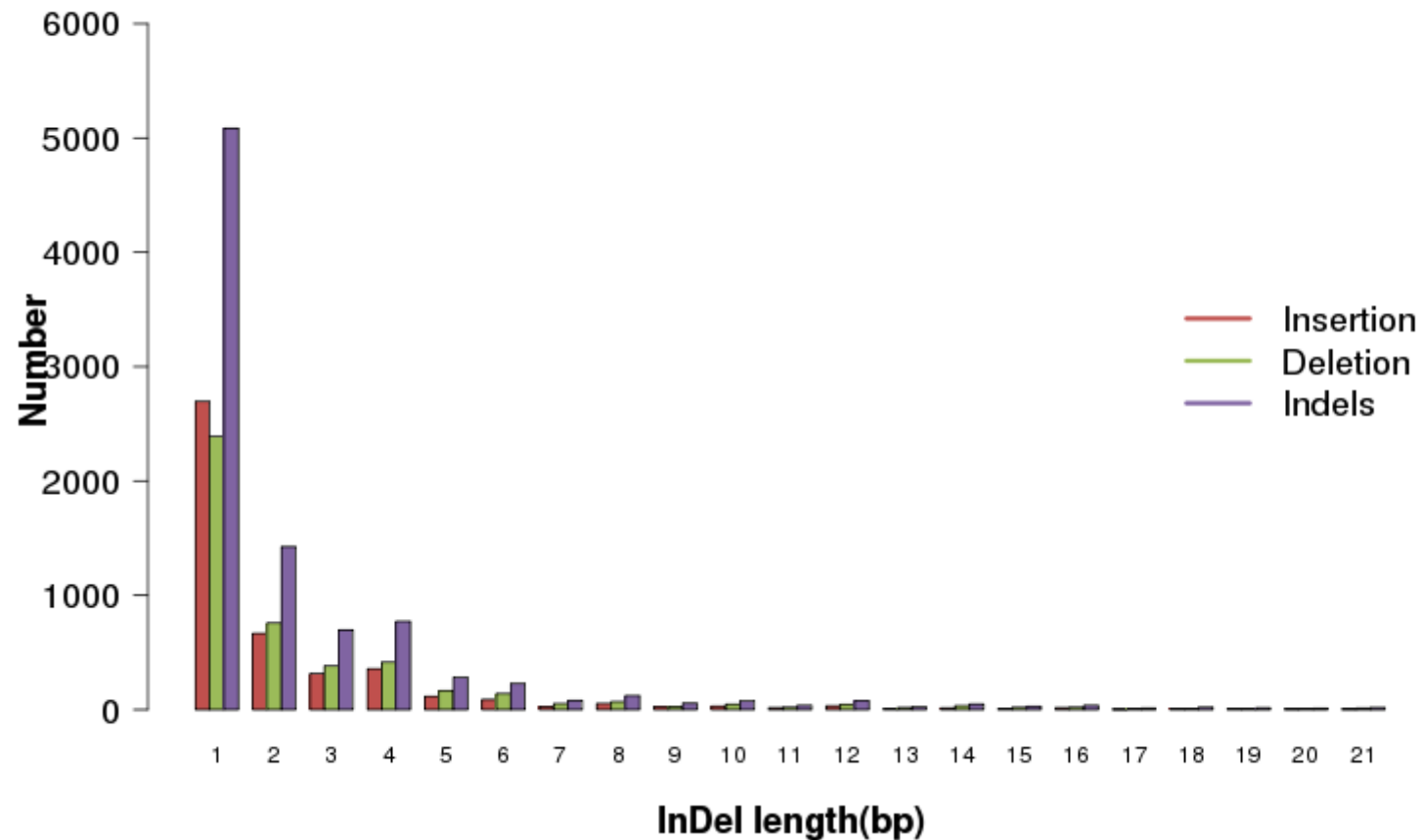


Figure 4.19 Family 2 Individual 2's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

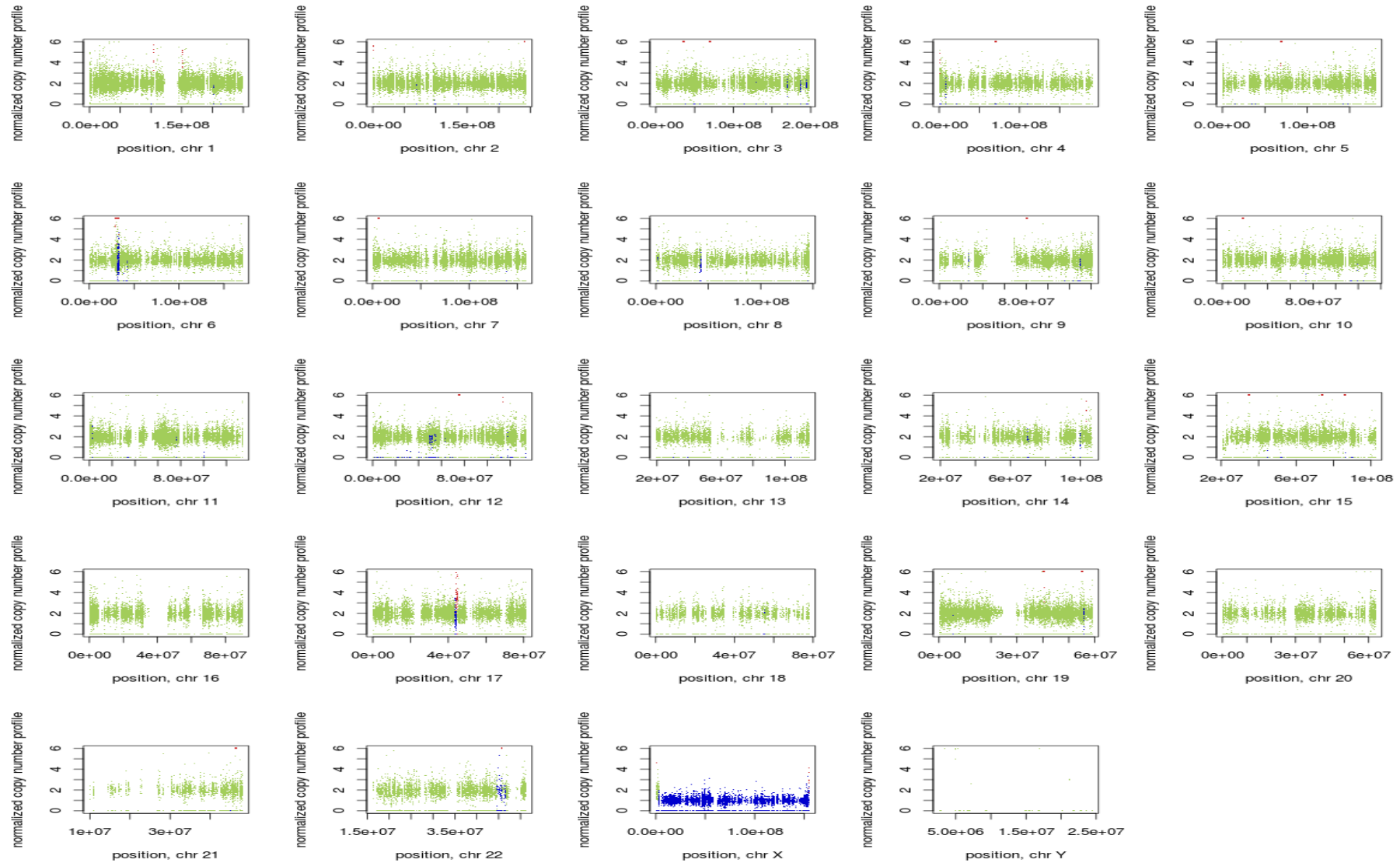


Figure 4.20 Family 2 Individual 2 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all blue chromosome X is expected when compared with a control female). None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Individual 3

Tables 4.20 and 4.21 summarises the SNV and indels detected and called for individual 3. Figures 4.21 and 4.22 depict the indel and CNV distribution throughout the genome as captured by WES platform.

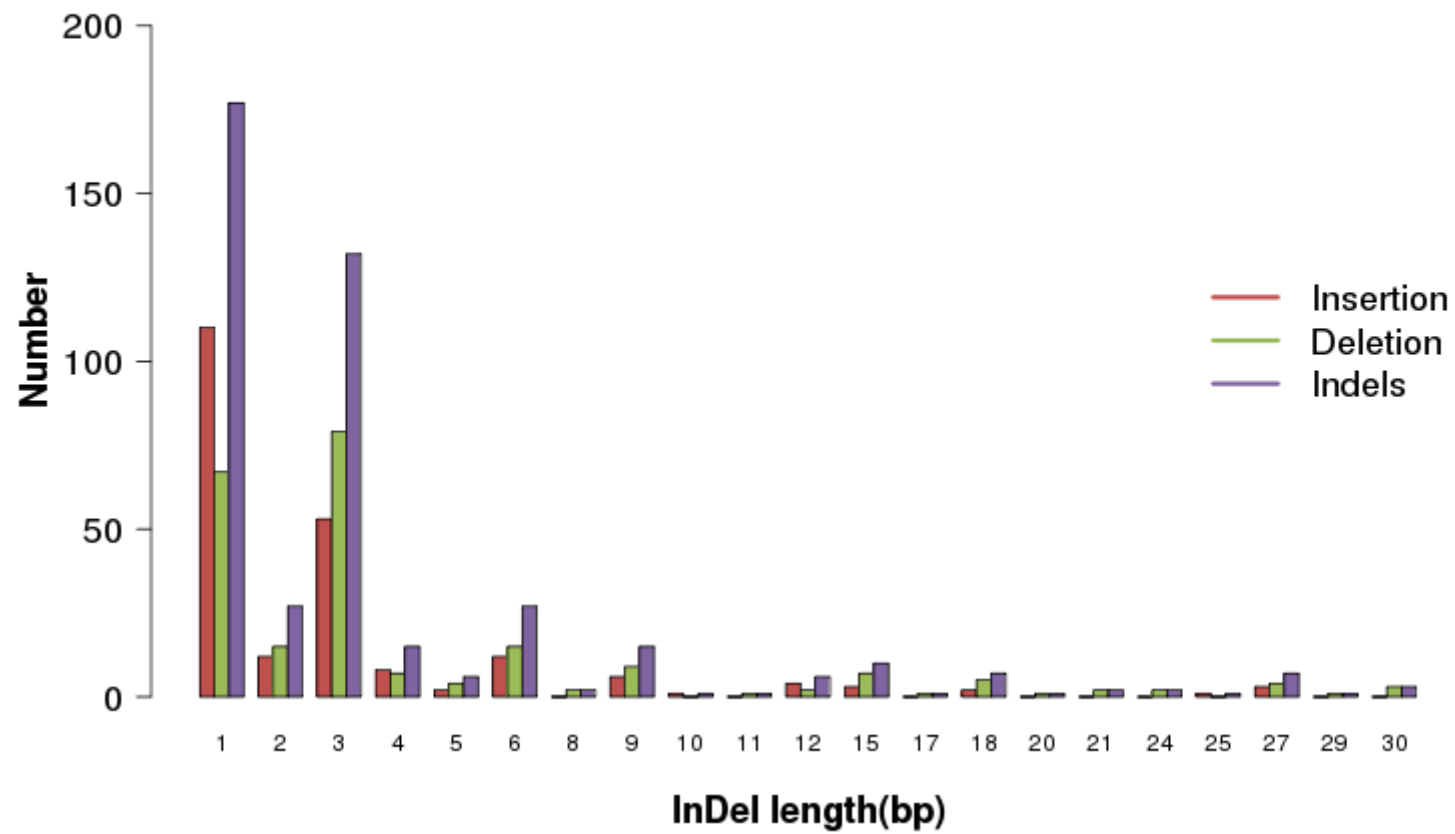
Categories	Value	Categories	Value
Total	79589	Splicing	60
1000genome and dbsnp135	75075	NcRNA	2642
1000genome specific	209	UTR5	931
dbSNP135 specific	2462	UTR5 and UTR3	2
dbSNP rate	97.42%	UTR3	2566
Novel	1843	Intronic	48565
Hom	33850	Upstream	602
Het	45739	Upstream and downstream	53
Synonymous	10039	Downstream	404
Missense	8880	Intergenic	4757
Stopgain	63	SIFT	1098
Stoploss	25	Ti/Tv	2.4250
Exonic	18755	dbSNP Ti/Tv	2.4343
Exonic and splicing	252	Novel Ti/Tv	2.1027

Table 4.20 Family 2 Individual 3's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9125	Stopgain	4
1000genome and dbsnp135	4493	Stoploss	1
1000genome specific	973	Exonic	403
dbSNP135 specific	1767	Exonic and splicing	5
dbSNP rate	68.60%	Splicing	38
Novel	1892	NcRNA	300
Hom	4163	UTR5	91
Het	4962	UTR5 and UTR3	0
Frameshift Insertion	107	UTR3	383
Non-frameshift Insertion	81	Intronic	7295
Frameshift Deletion	91	Upstream	77
Non-frameshift Deletion	124	Upstream and downstream	4
Frameshift block substitution	0	Downstream	51
Non-frameshift block substitution	0	Intergenic	478

Table 4.21 Family 2 Individual 3's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.20 above for a comparison.

InDel length distribution (CDS)



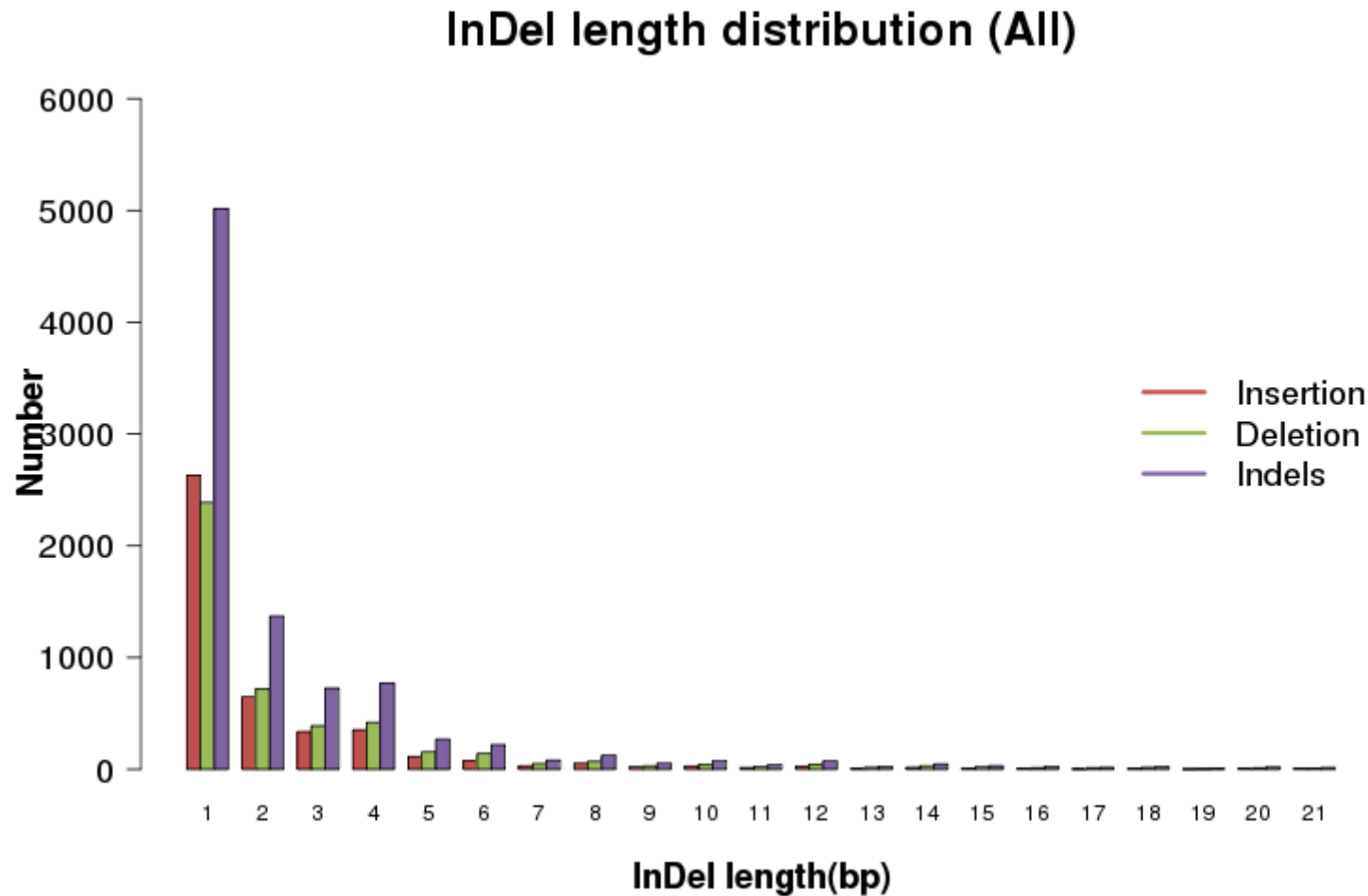


Figure 4.21 Family 2 Individual 3's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)'.

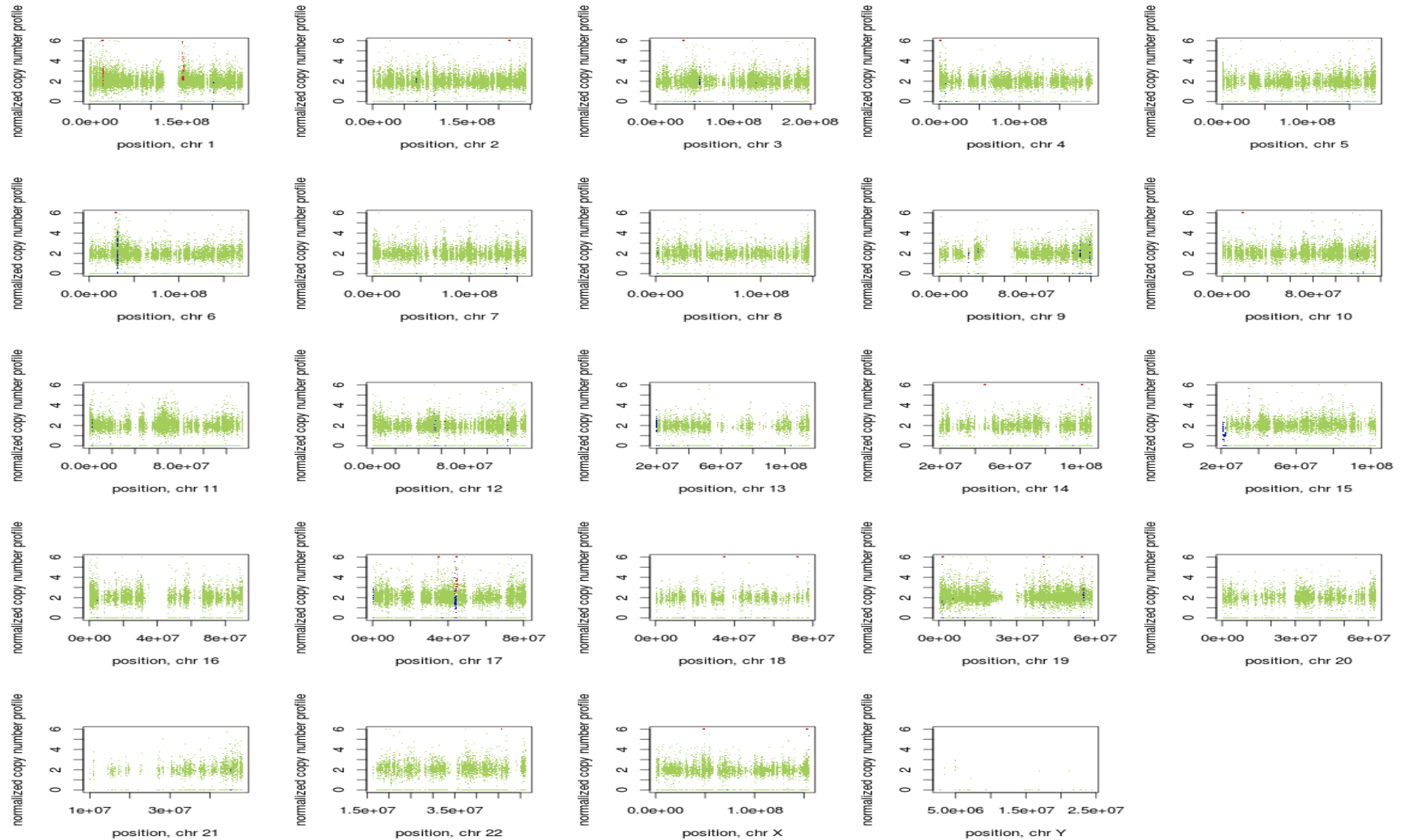


Figure 4.22 Family 2 Individual 3 CNV distribution. NB: Image may have been edited for confidentiality purposes. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels. None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Analysis of whole-exome, LRoH and candidate genes

Initially, all LRoH regions identified by Plink [294] and our custom Python script were reviewed in IGV [284] to check whether any known PCD gene (or reported linkage region) resides within the regions (Figures 4.23 to 4.25); and for these three siblings, matching LRoHs were also checked manually to see if they were shared by all three of them. None of the LRoHs was observed to be shared by all three siblings (i.e. individuals 1, 2 and 3). Then several filters (e.g. MAF, consequence of variant) were applied systematically on all mutations to single out any potentially causal ones. All remaining Φ mutations which were in a homozygous state were analysed separately in the two candidate lists. A final analysis was carried out on all the mutations outside of the two candidate lists.

There were only 8 homozygous Φ mutations in list 1 (i.e. known genes) which were shared by all three siblings; and another 197 homozygous Φ variants were identified by list 2 (i.e. all suspected ciliome genes). However all the mutations in list 1, and all except one in list 2 were common (i.e. with MAF = >1%); this was a missense mutation in the *CCT4* gene (p.R483Q). The mutation was not present in dbSNP, our internal database or in 1000GP, and was present in EVS with a total MAF of 0.0769% (0.093% in European Americans and 0.0454% in African Americans) and the ExAC database (86 heterozygotes in 121,330). However none of the individuals were homozygous for the allele. Assuming the individuals in EVS follow a Hardy-Weinberg equilibrium, then the probability of observing a homozygote (i.e. $q^2 = 0.000769^2$) for the mutation would be one in two million which could be the reason behind why the mutation was not identified before; and here we observe three siblings who are homozygotes for the mutation. The mutation does not fall into a shared LRoH but on close inspection, there is a ~1.5Mbp haplotype shared across the *CCT4* gene (conservative figure due to gaps in WES). It resides in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 1539.2 [310]. The missense mutation was predicted to be 'deleterious' by SIFT (score: 0.04) [243] and to an extent by (unweighted) FATHMM (-0.72, negative values indicating 'deleteriousness') [242] and Polyphen-2 (0.719, 'possibly damaging') [244]. A final

analysis was carried out on all the mutations outside of the two candidate lists and just 1 additional mutation was identified which was a frameshift mutation found in a homozygous state in 4 of our 13 internal controls (of Arabic ancestry), indicating non-causality. The additional CNV analysis yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

However when the region capturing the variant in the parents were sequenced using Sanger sequencing (using primers in Table 4.3), the variant was found to be in a homozygous state in the mother.

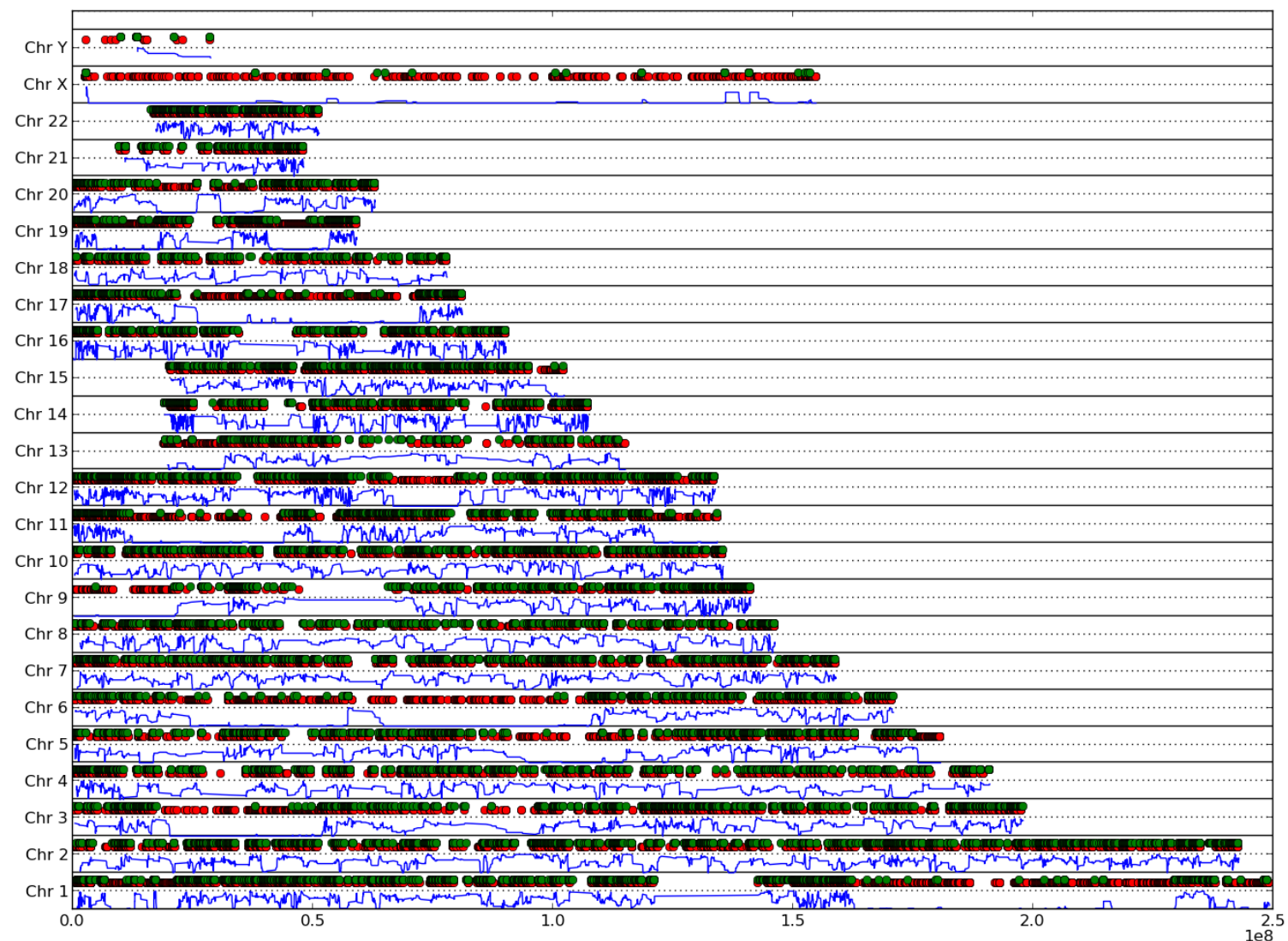


Figure 4.23 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 3 has an autozygous region ~20Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter. NB: These images may have been edited to ensure confidentiality/anonymity of the participants. Some LRoHs may also have been shortened/extended for the same reason.

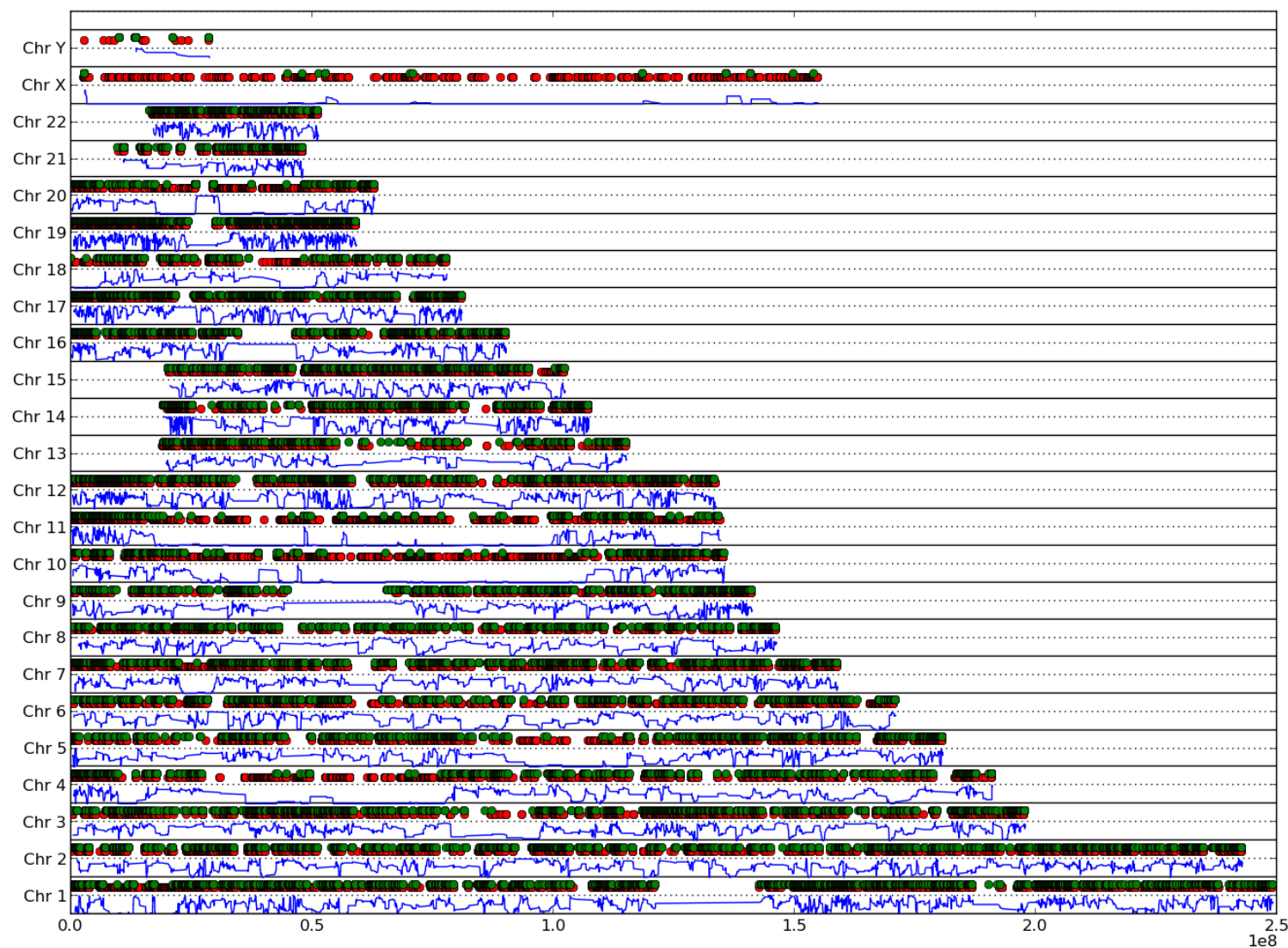


Figure 4.24 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 2. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 4 has an autozygous region ~20Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

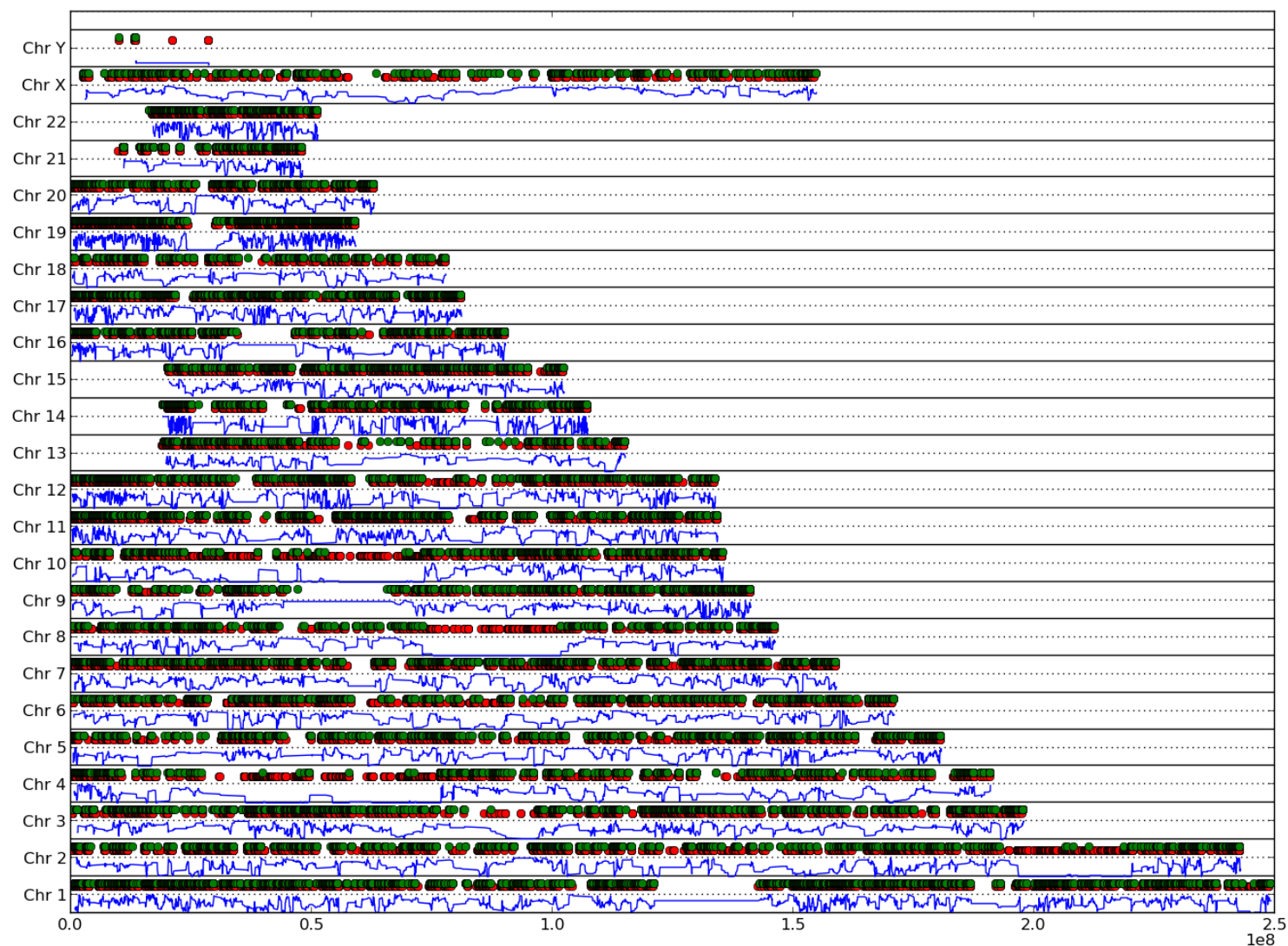


Figure 4.25 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 3. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 8 has an autozygous region ~25Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

4.5.4. Family 3

This section will present the results for family 3 and describe the findings:

No of whole-exome sequenced participants: 1

No of other family members with DNA available (incl. parents): 0

WES Data statistics

Total captured region was 115,650,744 base pairs (50,156,497 bases on target and 65,494,247 bases near target). Coverage of target and flanking regions was 97.3% and 89.4% respectively. The average sequencing depth on target was 61.05 and the fraction of target covered with at least 20 and 10 reads was 69.5% and 81.8% respectively (and >4 read depth = 91.3%). There were a total of 51,375,545 reads with a mapping rate of 99.51%.

Table 4.22 summarises the data produced by the WES platform for individual 1.

Type	Raw data	Clean data
Number of Reads	55918462	53282498
Data Size	5591846200	5328249800
N of fq1	31613	28844
N of fq2	69923	21087
GC(%) of fq1	42.58~42.65	42.43~42.51
GC(%) of fq2	42.75~42.81	42.56~42.63
Q20(%) of fq1	95.88~96.11	96.76~96.94
Q20(%) of fq2	92.39~92.60	95.35~95.50
Q30(%) of fq1	88.93~89.44	90.04~90.50
Q30(%) of fq2	85.17~85.61	88.07~88.45
Discard Reads related to N	3434	
Discard Reads related to low quality	2404360	
Discard Reads related to Adapter	228170	
Clean data/Raw data	95.29%	

Table 4.22 Family 3 Individual 1's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values. Clean to raw data ration was also high, indicating the quality of the data produced from the sequencing platform was high.

Read alignment statistics

Table 4.23 summarises the quality of alignment of the reads produced by WES platform to the reference genome for individual 1.

Exome Capture Statistics	Value	Exome Capture Statistics	Value
Initial bases on target	51543125	Average sequencing depth near target	15.76
Initial bases near target	73252334	Mismatch rate in target region	0.33%
Initial bases on or near target	124795459	Mismatch rate in all effective sequence	0.29%
Total effective reads	51375545	Base covered on target	50156497
Total effective yield(Mb)	5060.93	Coverage of target region	97.3%
Average read length(bp)	98.51	Base covered near target	65494247
Effective sequences on target(Mb)	3146.54	Coverage of flanking region	89.4%
Effective sequences near target(Mb)	1154.39	Fraction of target covered with at least 20x	69.5%
Effective sequences on or near target(Mb)	4300.93	Fraction of target covered with at least 10x	81.8%
Number of reads uniquely mapped to target	34841272	Fraction of target covered with at least 4x	91.3%
Number of reads uniquely mapped to genome	46243369	Fraction of flanking region covered with at least 20x	24.9%
Fraction of effective bases on target	62.2%	Fraction of flanking region covered with at least 10x	42.9%
Fraction of uniquely mapped on target	75.3%	Fraction of flanking region covered with at least 4x	67.1%
Fraction of effective bases on or near target	85.0%	Mapping rate	99.51%
Average sequencing depth on target	61.05	Duplicate rate	3.10%

Table 4.23 Family 3 Individual 1's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference genome.

Variant calling statistics

Tables 4.24 and 4.25 summarises the SNV and indels detected and called for individual 1 respectively. Figures 4.26 and 4.27 depict the indel and CNV distribution throughout the genome as captured by WES platform.

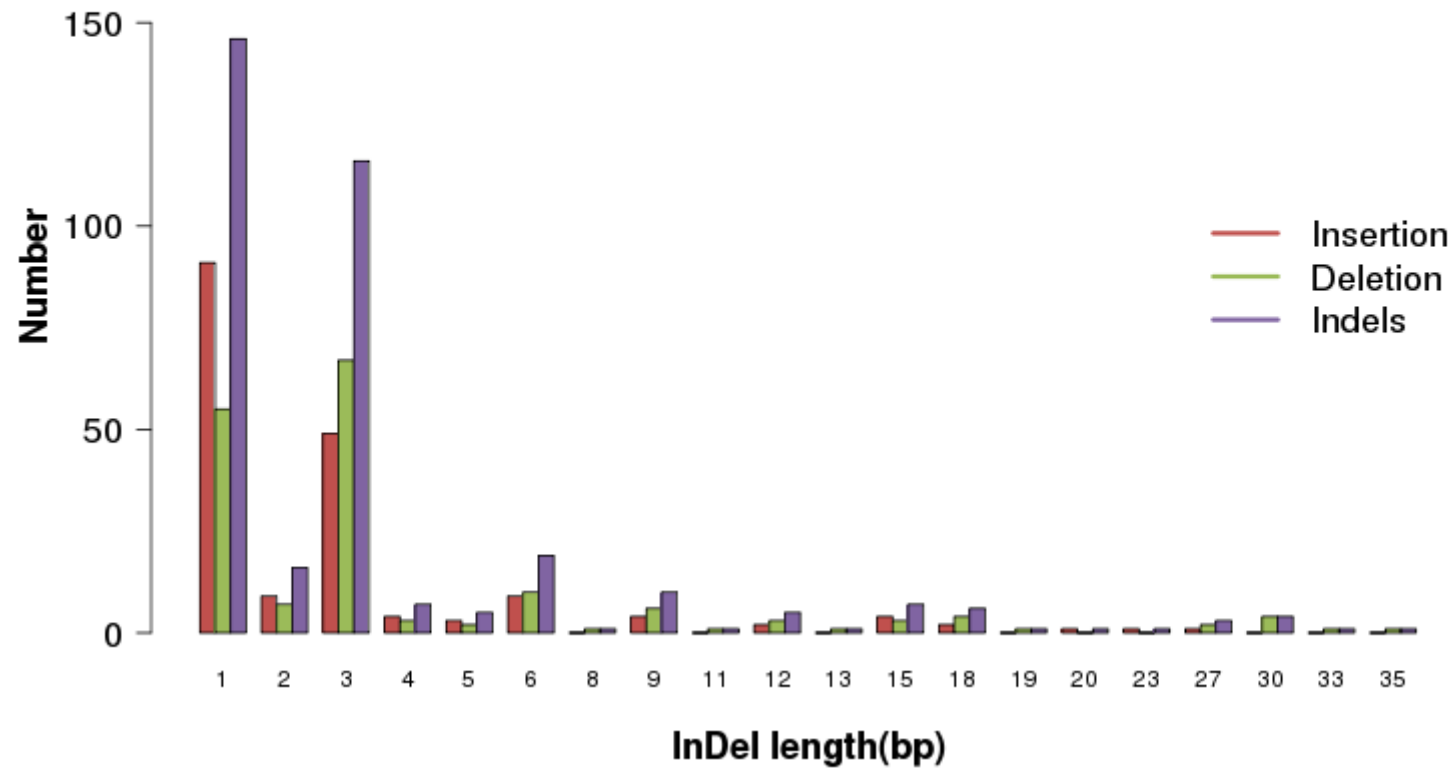
Categories	Value	Categories	Value
Total	70888	Splicing	51
1000genome and dbsnp135	67087	NcRNA	2444
1000genome specific	202	UTR5	775
dbSNP135 specific	2127	UTR5 and UTR3	3
dbSNP rate	97.64%	UTR3	2164
Novel	1472	Intronic	43346
Hom	33260	Upstream	571
Het	37628	Upstream and downstream	39
Synonymous	8916	Downstream	355
Missense	7730	Intergenic	4411
Stopgain	55	SIFT	922
Stoploss	28	Ti/Tv	2.3973
Exonic	16509	dbSNP Ti/Tv	2.4059
Exonic and splicing	220	Novel Ti/Tv	2.0924

Table 4.24 Family 3 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	8578	Stopgain	1
1000genome and dbsnp135	4296	Stoploss	0
1000genome specific	880	Exonic	320
dbSNP135 specific	1737	Exonic and splicing	2
dbSNP rate	70.33%	Splicing	30
Novel	1665	NcRNA	297
Hom	4056	UTR5	74
Het	4522	UTR5 and UTR3	0
Frameshift Insertion	86	UTR3	343
Non-frameshift Insertion	69	Intronic	6938
Frameshift Deletion	68	Upstream	70
Non-frameshift Deletion	98	Upstream and downstream	5
Frameshift block substitution	0	Downstream	59
Non-frameshift block substitution	0	Intergenic	440

Table 4.25 Family 3 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.24 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

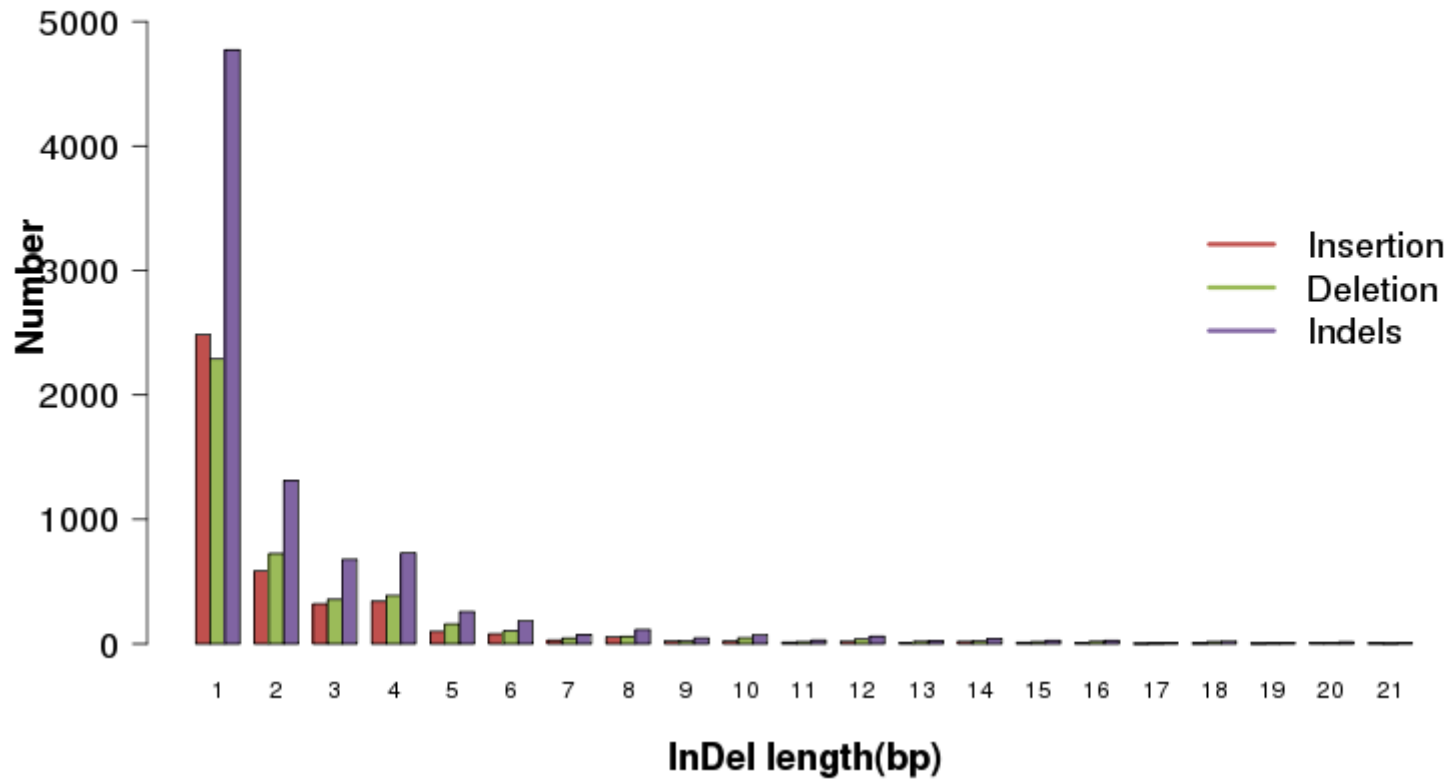


Figure 4.26 Family 3 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

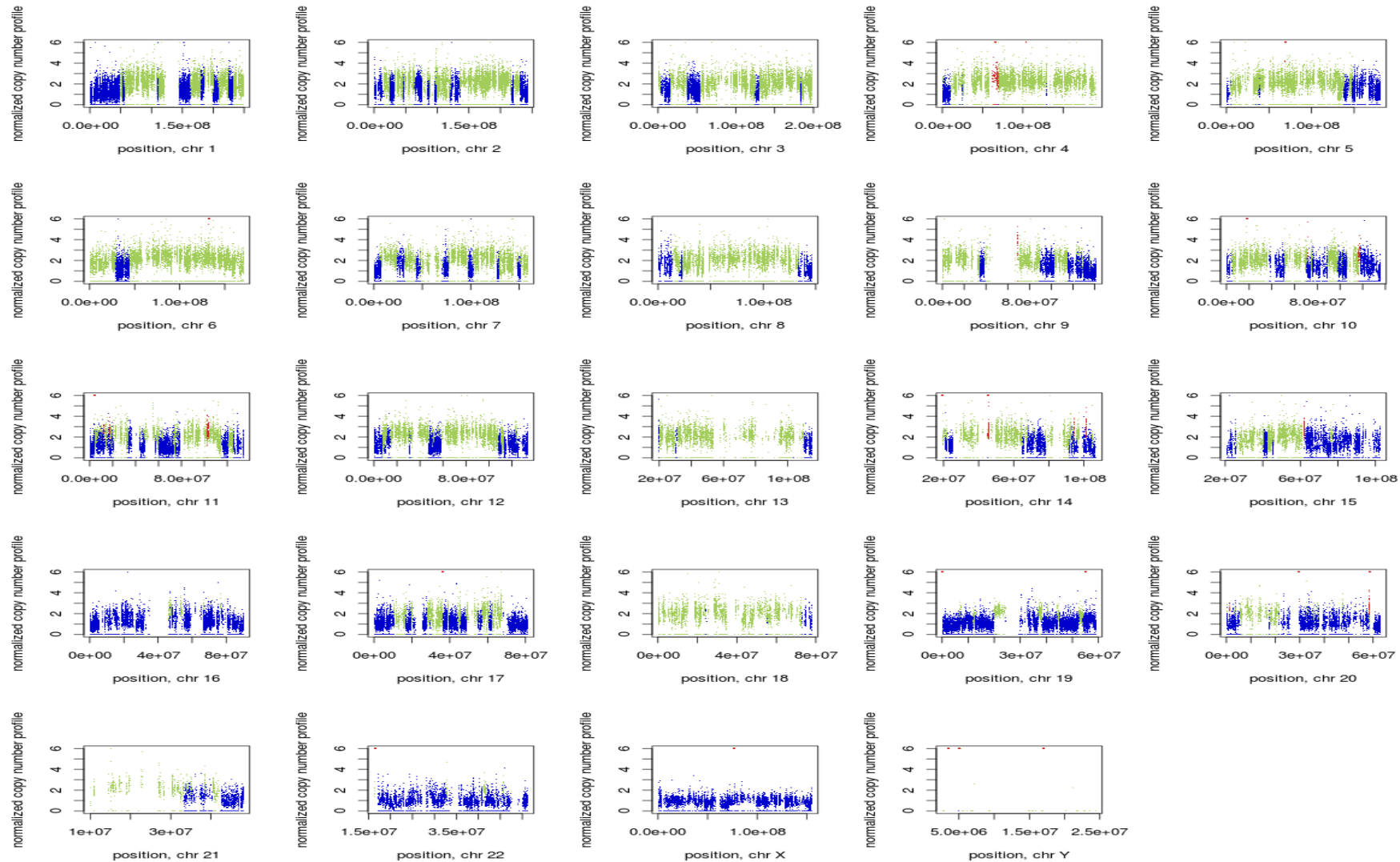


Figure 4.27 Family 3 Individual 1's CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all blue chromosome X is expected when compared with a control female). This individual has overall less depth than the control, therefore blues are expected throughout the genome. However none of the deviations that overlapped with an already known PCD causal gene were fully affected. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Analysis of whole-exome, LRoH and candidate genes

Analysis of list 1 returned 14 homozygous Φ variants; and list 2 returned an additional 318 Φ mutations. As aforementioned, one of the LRoHs of patient 3 (~23Mb long) contained a known PCD gene in *DNAH11* and a missense mutation (p.T4338M with a read depth of 101) was found within the gene (Figure 4.28). The variant was absent in 1000GP, dbSNP, EVS, ExAC and our internal database. However this mutation was predicted to be 'tolerated' by all four of the mutation prediction tools we used in this analysis and was not predicted to be 'conserved' by GERP. Analysis of list 1 showed that the patient had one more mutation in known genes which was a frameshifting deletion (p.G734fs with read depth of 19) in the *HEATR2* gene. The mutation was absent in all 4 databases we used in the analysis. However, the mutation was not in an LRoH and was also not predicted to be 'conserved' by GERP [310]. Another potentially causal mutation, a stop gain (p.E328* with read depth of 11) in the *LRRC48* gene was observed in list 2. It was not present in any of the databases analysed. This mutation also fell in an LRoH (~9Mb long) and in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 58.6 [310]. All other mutations were observed to be common in at least one of the 4 databases we used in this analysis.

Since patient 3 also had thrombocytopenia (see section 4.4.2), mutations in *ADAMTS13* were also specifically analysed [414]. However all variants in his *ADAMTS13* gene were common (>1%), indicating it is more likely an acquired form rather than the familial form. The additional CNV analysis yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

This analysis yielded two novel candidate variants for PCD which require follow up via further studies.

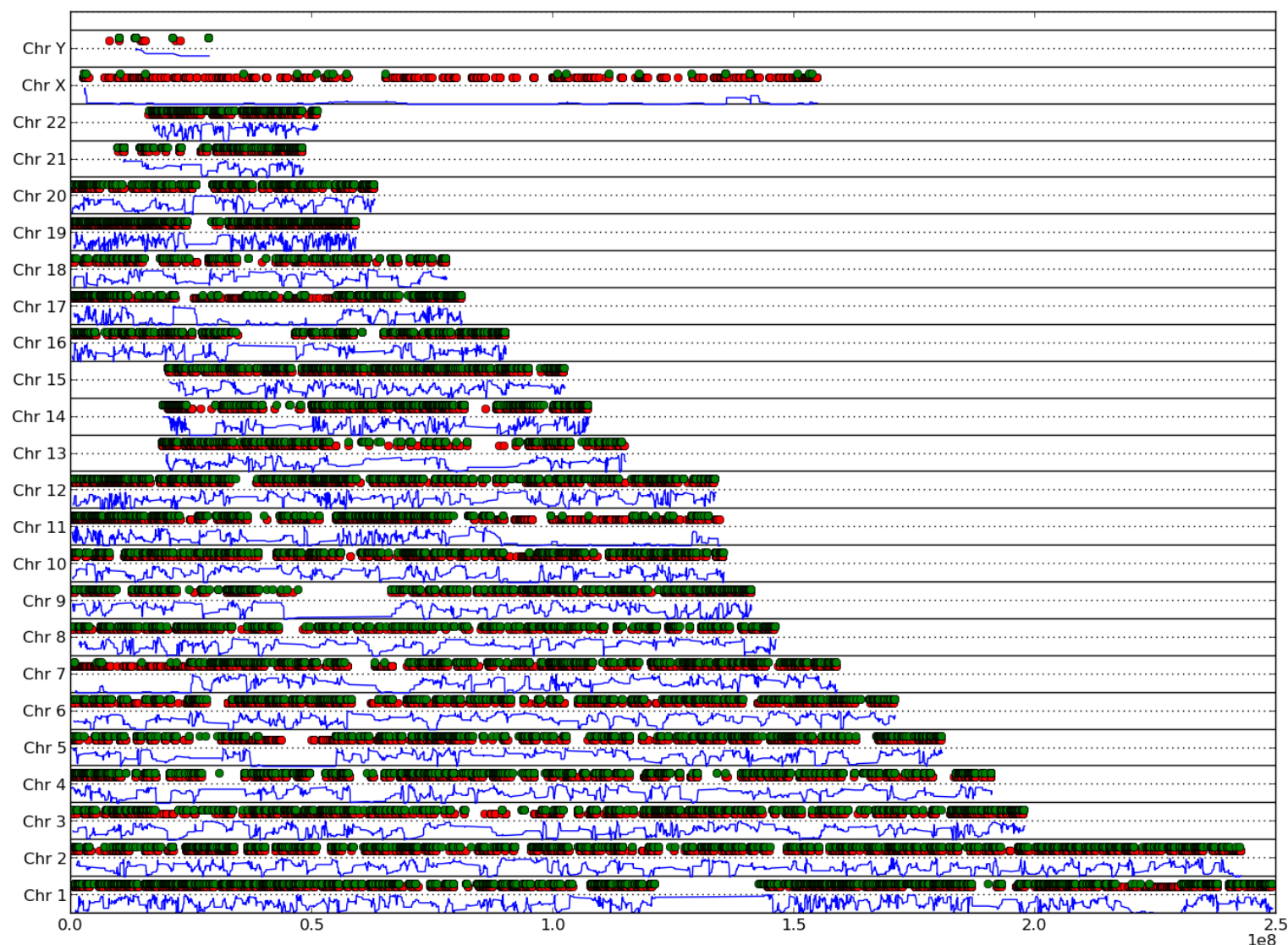


Figure 4.28 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 7 has an autozygous region ~27Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

4.5.5. Family 4

This section will present the results for family 4 and describe the findings:

No of whole-exome sequenced participants: 1

No of other family members with DNA available (incl. parents): 0

WES Data statistics

Total captured region was 118,981,699 base pairs (50,689,834 bases on target and 68,291,865 bases near target). Coverage of target and flanking regions was 98.3% and 93.2% respectively. The average sequencing depth on target was 60.55 and the fraction of target covered with at least 20 and 10 reads was 79.5% and 88.8% respectively (and >4 read depth = 94.8%). There were a total of 51,817,236 (high quality) reads with a mapping rate of 99.37%.

Table 4.26 summarises the data produced by the WES platform for individual 1.

Type	Raw data	Clean data
Number of Reads	57165868	54128906
Data Size	5716586800	5412890600
N of fq1	31631	28629
N of fq2	73627	22371
GC(%) of fq1	44.98~45.06	44.78~44.88
GC(%) of fq2	45.16~45.24	44.9~44.99
Q20(%) of fq1	95.43~95.68	96.48~96.67
Q20(%) of fq2	91.45~91.68	94.88~95.03
Q30(%) of fq1	87.99~88.54	89.32~89.81
Q30(%) of fq2	83.66~84.12	87.00~87.39
Discard Reads related to N	3588	
Discard Reads related to low quality	2842518	
Discard Reads related to Adapter	190856	
Clean data/Raw data	94.69%	

Table 4.26 Family 4 Individual 1's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values. Clean to raw data ration was also high, indicating the quality of the data produced from the sequencing platform was high.

Read alignment statistics

Table 4.27 summarises the quality of alignment of the reads produced by WES platform to the reference genome for individual 1.

Exome Capture Statistics	Value	Exome Capture Statistics	Value
Initial bases on target	51543125	Average sequencing depth near target	15.95
Initial bases near target	73252334	Mismatch rate in target region	0.33%
Initial bases on or near target	124795459	Mismatch rate in all effective sequence	0.30%
Total effective reads	51817236	Base covered on target	50689834
Total effective yield(Mb)	5094.20	Coverage of target region	98.3%
Average read length(bp)	98.31	Base covered near target	68291865
Effective sequences on target(Mb)	3120.77	Coverage of flanking region	93.2%
Effective sequences near target(Mb)	1168.28	Fraction of target covered with at least 20x	79.5%
Effective sequences on or near target(Mb)	4289.05	Fraction of target covered with at least 10x	88.8%
Number of reads uniquely mapped to target	34196278	Fraction of target covered with at least 4x	94.8%
Number of reads uniquely mapped to genome	46130578	Fraction of flanking region covered with at least 20x	26.3%
Fraction of effective bases on target	61.3%	Fraction of flanking region covered with at least 10x	47.6%
Fraction of uniquely mapped on target	74.1%	Fraction of flanking region covered with at least 4x	73.6%
Fraction of effective bases on or near target	84.2%	Mapping rate	99.37%
Average sequencing depth on target	60.55	Duplicate rate	3.66%

Table 4.27 Family 4 Individual 1's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference genome.

Variant calling statistics

Tables 4.28 and 4.29 summarises the SNV and indels detected and called for individual 1 respectively. Figures 4.29 and 4.30 depict the indel and CNV distribution throughout the genome as captured by WES platform.

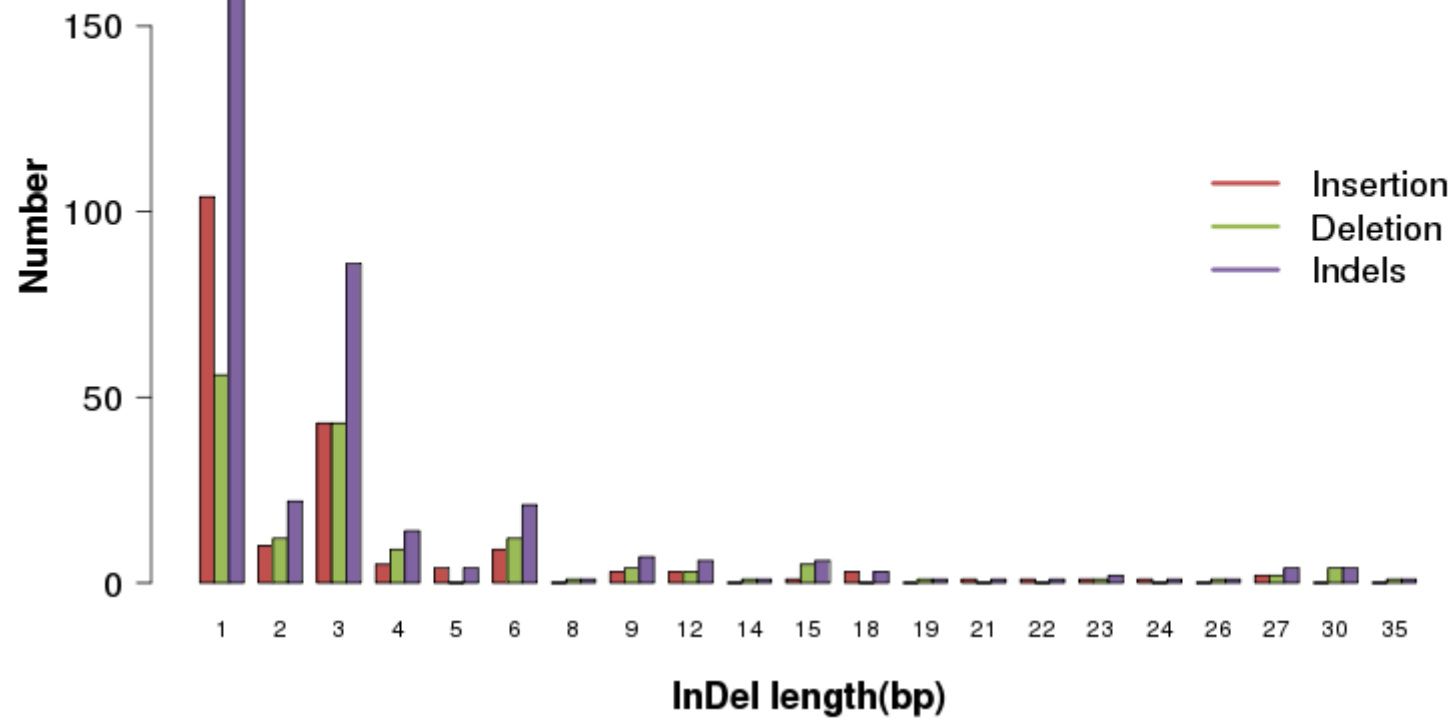
Categories	Value	Categories	Value
Total	79023	Splicing	59
1000genome and dbsnp135	74823	NcRNA	2732
1000genome specific	191	UTR5	930
dbSNP135 specific	2360	UTR5 and UTR3	5
dbSNP rate	97.67%	UTR3	2529
Novel	1649	Intronic	48769
Hom	35455	Upstream	610
Het	43568	Upstream and downstream	39
Synonymous	9797	Downstream	420
Missense	8522	Intergenic	4514
Stopgain	72	SIFT	1062
Stoploss	25	Ti/Tv	2.4203
Exonic	18161	dbSNP Ti/Tv	2.4317
Exonic and splicing	255	Novel Ti/Tv	2.0594

Table 4.28 Family 4 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9081	Stopgain	3
1000genome and dbsnp135	4594	Stoploss	0
1000genome specific	898	Exonic	396
dbSNP135 specific	1856	Exonic and splicing	2
dbSNP rate	71.03%	Splicing	35
Novel	1733	NcRNA	333
Hom	4281	UTR5	91
Het	4800	UTR5 and UTR3	0
Frameshift Insertion	105	UTR3	383
Non-frameshift Insertion	90	Intronic	7253
Frameshift Deletion	82	Upstream	66
Non-frameshift Deletion	118	Upstream and downstream	5
Frameshift block substitution	0	Downstream	68
Non-frameshift block substitution	0	Intergenic	449

Table 4.29 Family 4 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.28 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

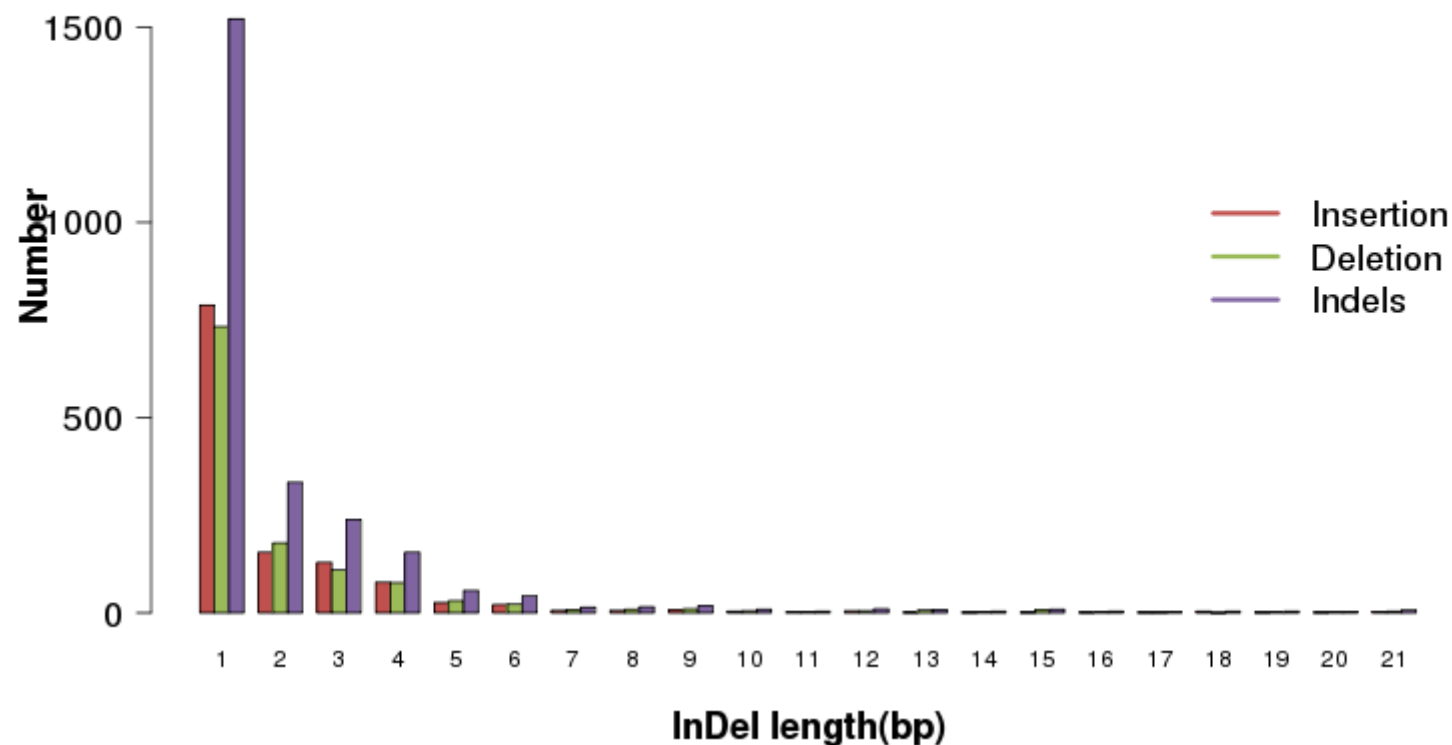


Figure 4.29 Family 4 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

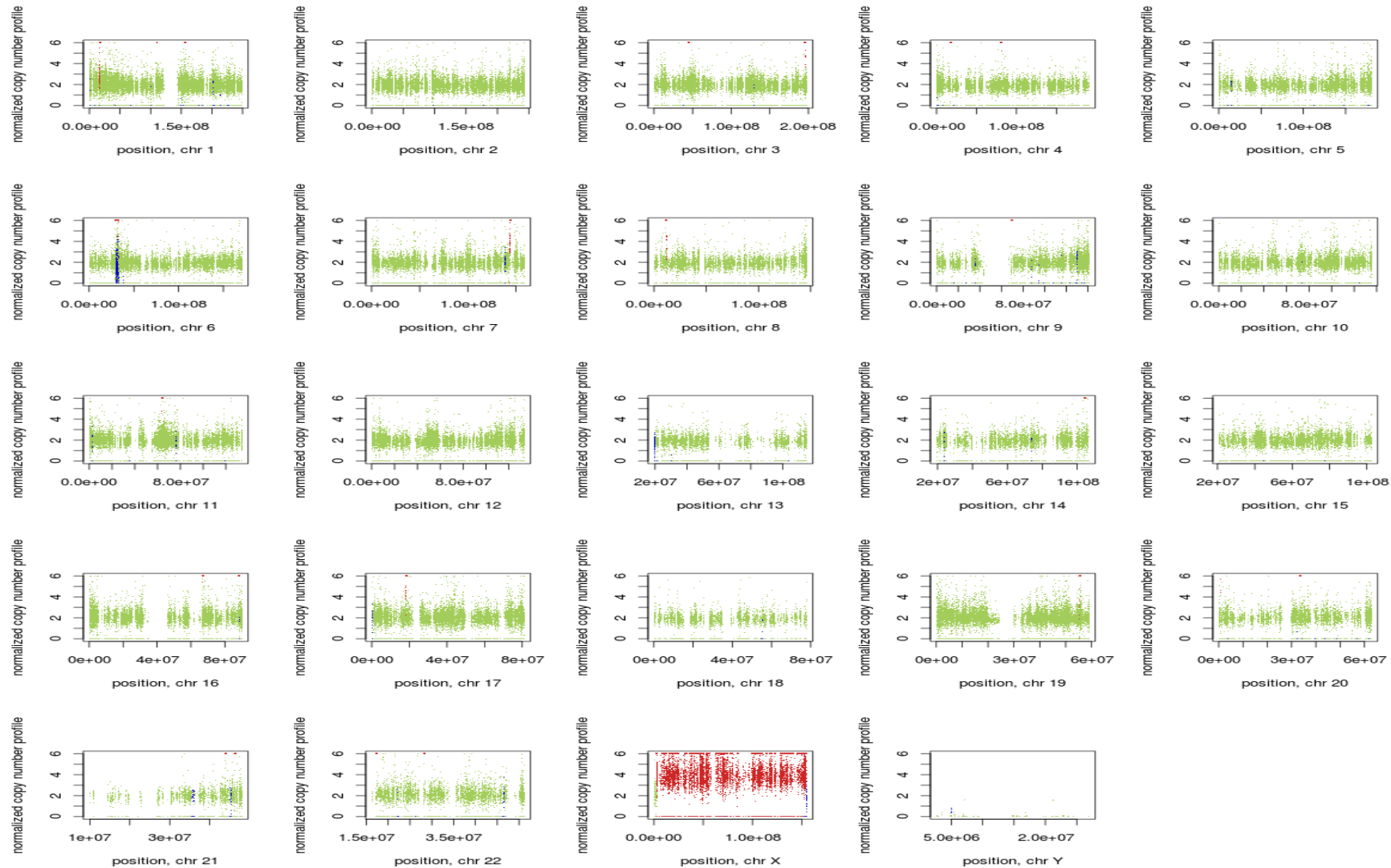


Figure 4.30 Family 4 Individual 1 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all red chromosome X is expected when a female is compared with a control male). None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Analysis of whole-exome, LRoH and candidate genes

Analysing patient 4's LRoHs mapped one of them (~3.3Mb long) to a known PCD causal gene, *DNAAF3* (Figure 4.31). Analysing this region yielded a nonsense mutation in *DNAAF3* itself (p.R136* with a read depth of 35). The stop gain falls near the centre of the protein (which is 541 amino acids long) thus the mutant transcript being a target for NMD would be an expected consequence [415]. Also it is found to reside in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 272 [310]. The mutation was absent in 1000GP, dbSNP, EVS, ExAC and our internal database. However, the variant has already been identified by a previous study carried out by Mitchison *et al* which raises the possibility of this variant being a founder mutation (NB: the previously studied subject by Mitchison *et al* was also of Arabic ancestry) [34]. The additional CNV analysis yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

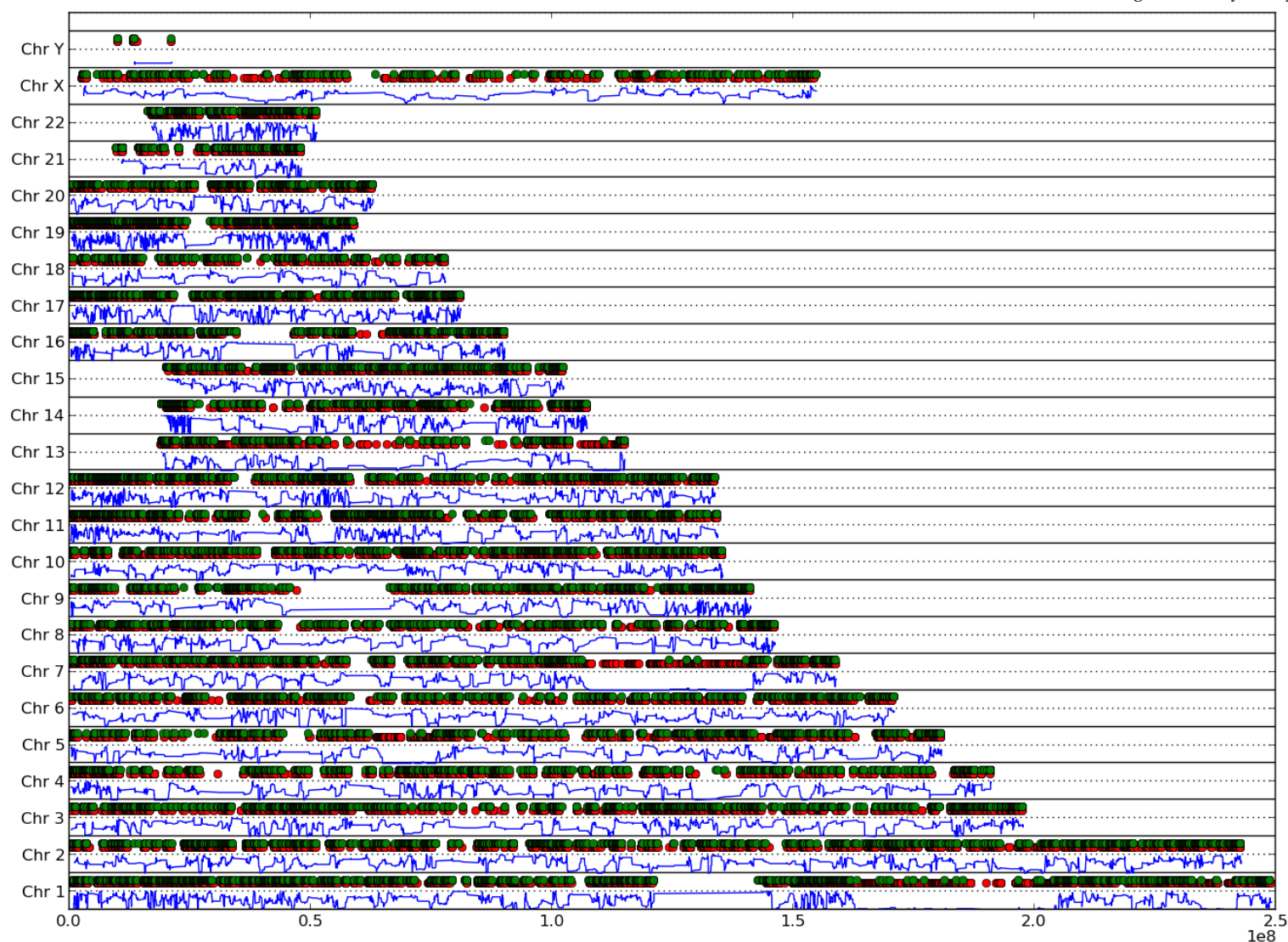


Figure 4.31 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 1 has an autozygous region ~30Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter. NB: These images may have been edited to ensure confidentiality/anonymity of the participants. Some LRoHs may also have been shortened/extended for the same reason.

Modelling mutation effect on protein structure

Uploading the amino acid sequences stated in section 4.4.9 to the Robetta server yielded 5 different models for both the mutant and the wild type DNAAF3 protein. Figures 4.32 and 4.33 below depict the first model for both. The affected residue is labelled in both.

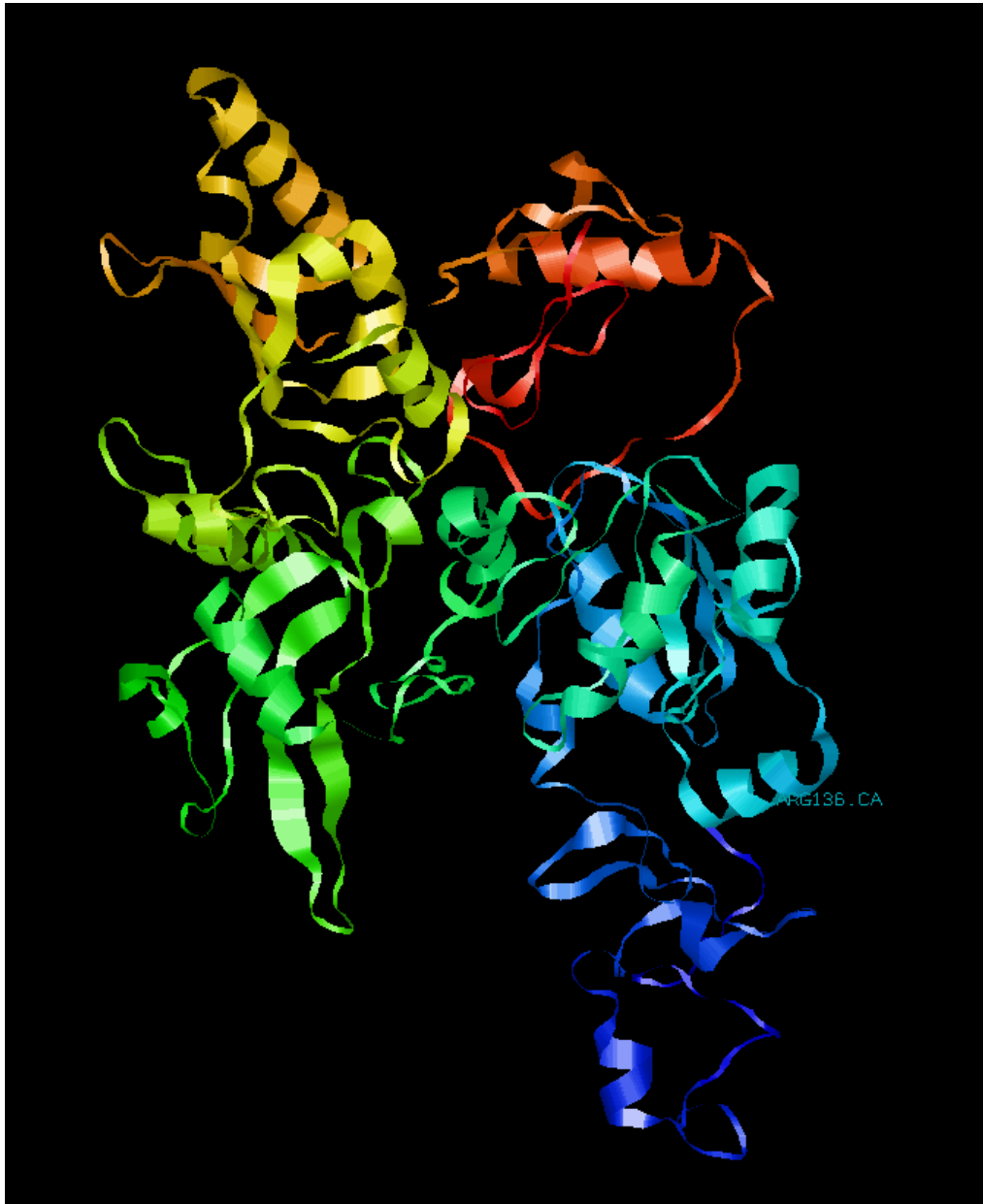


Figure 4.32 Protein structure of wild type DNAAF3 protein. Only the first (most likely) model is shown and the location of the mutation (p.R136*) is labelled. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.



Figure 4.33 Protein structure of mutant DNAAF3 protein. Only the first model is shown. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol. The mutation p.R136* cannot be presented as the protein product is truncated from the stopgain locus (see Figure 4.32). It is easy to observe that most of the protein domains have been removed, and therefore to comprehend how such a large structural change can affect the correct functioning of the protein. See discussion on section 4.6.2 for details.

4.5.6. Family 5

This section will present the results for family 5 and describe the findings:

No of whole-exome sequenced participants: 1

No of other family members with DNA available (incl. parents): 0

WES Data statistics

Total captured region was 120,679,925 base pairs (50,721,434 bases on target and 69,957,491 bases near target). Coverage of target and flanking regions was 98.4% and 95.5% respectively. The average sequencing depth on target was 60.45 and the fraction of target covered with at least 20 and 10 reads was 78.2% and 88.1% respectively (and >4 read depth = 94.6%). There were a total of 76,677,082 reads with a mapping rate of 98.7%.

Table 4.30 summarises the data produced by the WES platform for individual 1.

Type	Raw data	Clean data
Number of Reads	84836718	81089752
Data Size	8483671800	8108975200
N of fq1	143691	113940
N of fq2	165039	134757
GC(%) of fq1	43.06~43.07	42.93~42.93
GC(%) of fq2	43.25~43.25	43~43.02
Q20(%) of fq1	96.89~96.94	97.63~97.68
Q20(%) of fq2	93.95~93.96	96.64~96.66
Q30(%) of fq1	91.39~91.52	92.34~92.47
Q30(%) of fq2	88.24~88.25	90.93~90.93
Discard Reads related to N	4112	
Discard Reads related to low quality	3247610	
Discard Reads related to Adapter	495244	
Clean data/Raw data	95.58%	

Table 4.30 Family 5 Individual 1's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values. Clean to raw data ration was also high, indicating the quality of the data produced from the sequencing platform was high.

Read alignment statistics

Table 4.31 (next page) summarises the quality of alignment of the reads produced by WES platform to the reference genome for individual 1.

Exome Capture Statistics	Value
Initial bases on target	51543125
Initial bases near target	73252334
Initial bases on or near target	124795459
Total effective reads	76677082
Total effective yield(Mb)	7573.05
Average read length(bp)	98.77
Effective sequences on target(Mb)	3115.76
Effective sequences near target(Mb)	1238.26
Effective sequences on or near target(Mb)	4354.02
Number of reads uniquely mapped to target	34155140
Number of reads uniquely mapped to genome	69646871
Fraction of effective bases on target	41.1%
Fraction of uniquely mapped on target	49.0%
Fraction of effective bases on or near target	57.5%
Average sequencing depth on target	60.45

Exome Capture Statistics	Value
Average sequencing depth near target	16.90
Mismatch rate in target region	0.23%
Mismatch rate in all effective sequence	0.24%
Base covered on target	50721434
Coverage of target region	98.4%
Base covered near target	69957491
Coverage of flanking region	95.5%
Fraction of target covered with at least 20x	78.2%
Fraction of target covered with at least 10x	88.1%
Fraction of target covered with at least 4x	94.6%
Fraction of flanking region covered with at least 20x	27.7%
Fraction of flanking region covered with at least 10x	50.8%
Fraction of flanking region covered with at least 4x	78.6%
Mapping rate	98.70%
Duplicate rate	4.19%

Table 4.31 Family 5 Individual 1's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference genome.

Variant calling statistics

Tables 4.32 and 4.33 summarises the SNV and indels detected and called for individual 1 respectively. Figures 4.34 and 4.35 depict the indel and CNV distribution throughout the genome as captured by WES platform.

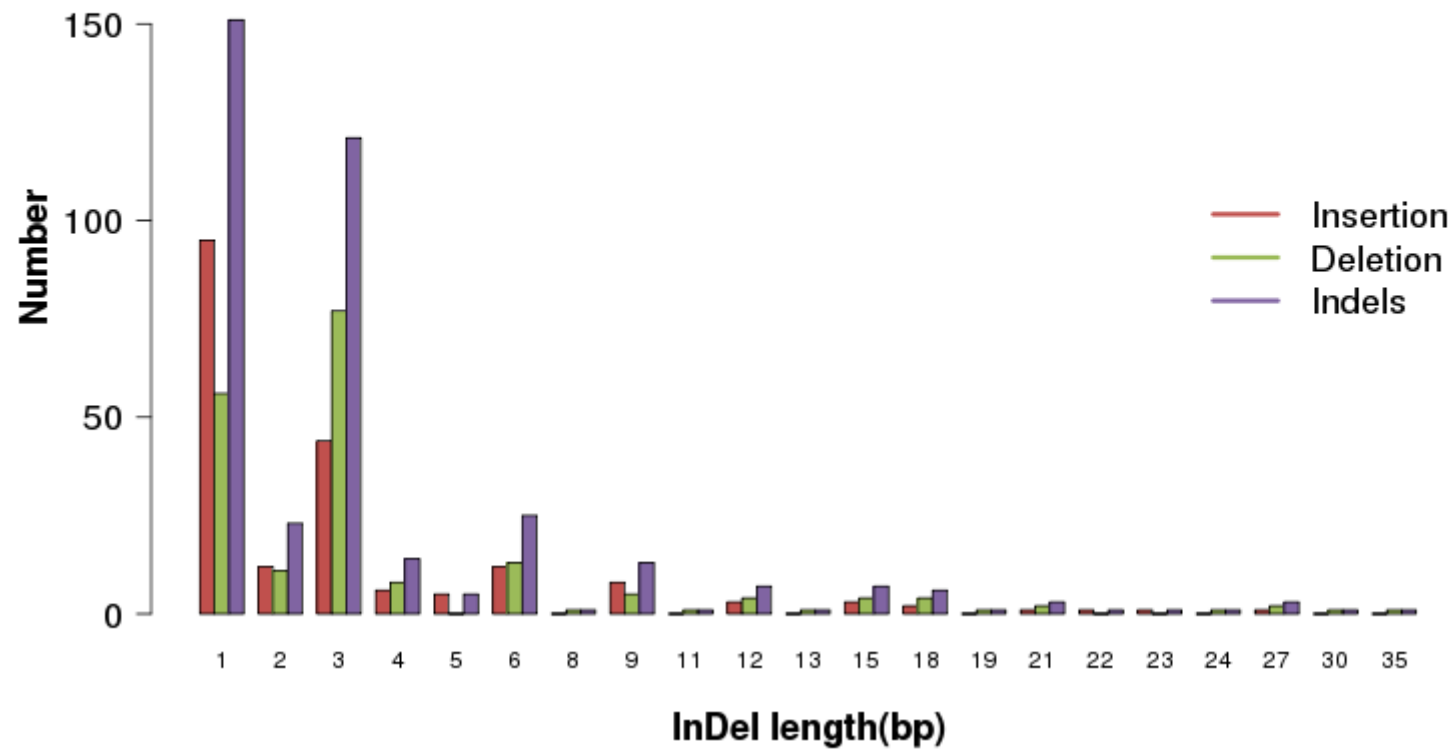
Categories	Value	Categories	Value
Total	84053	Splicing	55
1000genome and dbsnp135	79682	NcRNA	2700
1000genome specific	223	UTR5	939
dbSNP135 specific	2485	UTR5 and UTR3	4
dbSNP rate	97.76%	UTR3	2728
Novel	1663	Intronic	52651
Hom	36995	Upstream	700
Het	47058	Upstream and downstream	76
Synonymous	9887	Downstream	492
Missense	8747	Intergenic	4968
Stopgain	75	SIFT	1076
Stoploss	31	Ti/Tv	2.4277
Exonic	18502	dbSNP Ti/Tv	2.4391
Exonic and splicing	238	Novel Ti/Tv	2.0402

Table 4.32 Family 5 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9816	Stopgain	2
1000genome and dbsnp135	4894	Stoploss	0
1000genome specific	996	Exonic	350
dbSNP135 specific	2016	Exonic and splicing	5
dbSNP rate	70.40%	Splicing	32
Novel	1910	NcRNA	320
Hom	4442	UTR5	90
Het	5374	UTR5 and UTR3	0
Frameshift Insertion	97	UTR3	408
Non-frameshift Insertion	72	Intronic	7953
Frameshift Deletion	77	Upstream	75
Non-frameshift Deletion	107	Upstream and downstream	8
Frameshift block substitution	0	Downstream	72
Non-frameshift block substitution	0	Intergenic	503

Table 4.33 Family 5 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.32 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

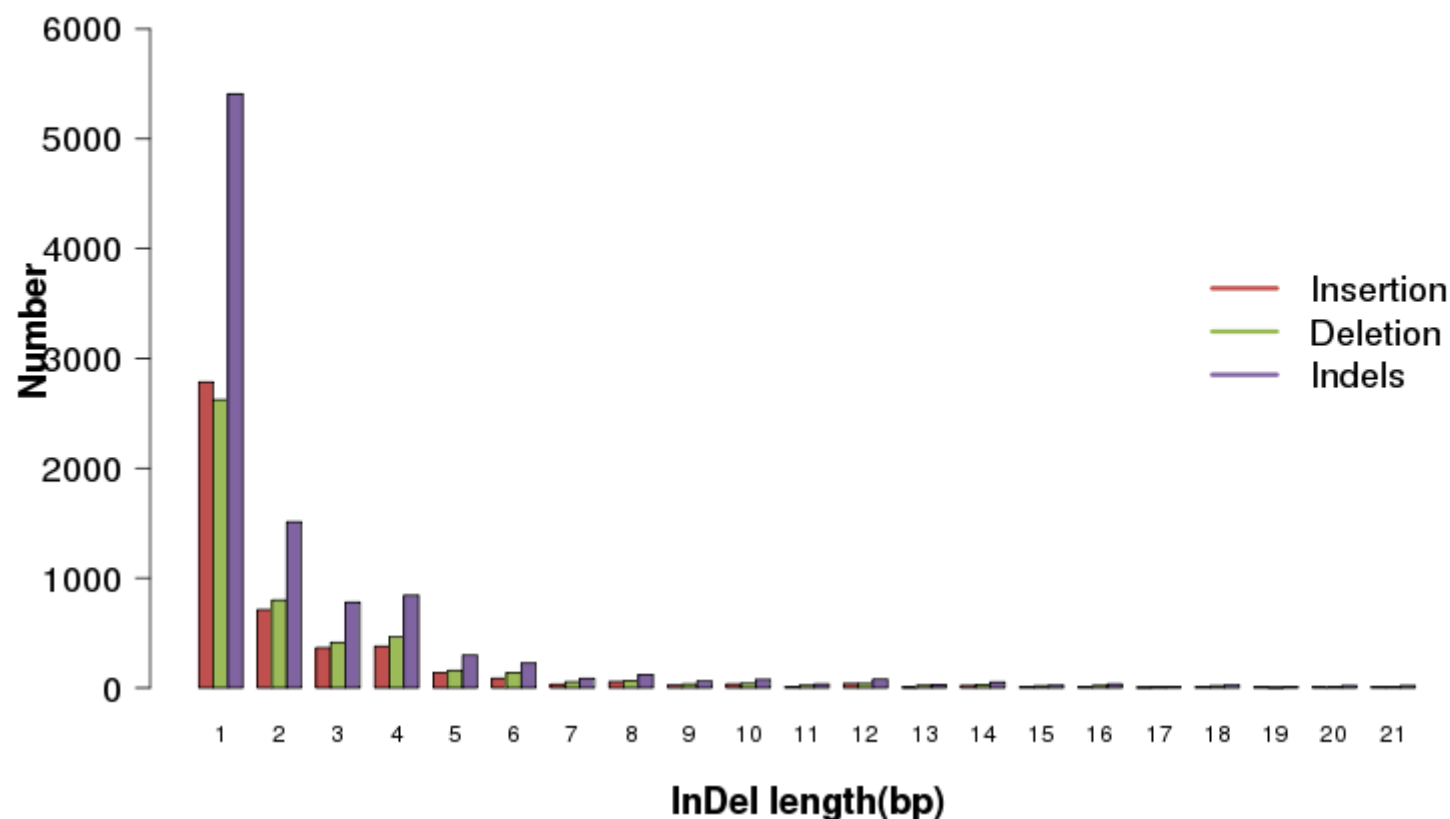


Figure 4.34 Family 5 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

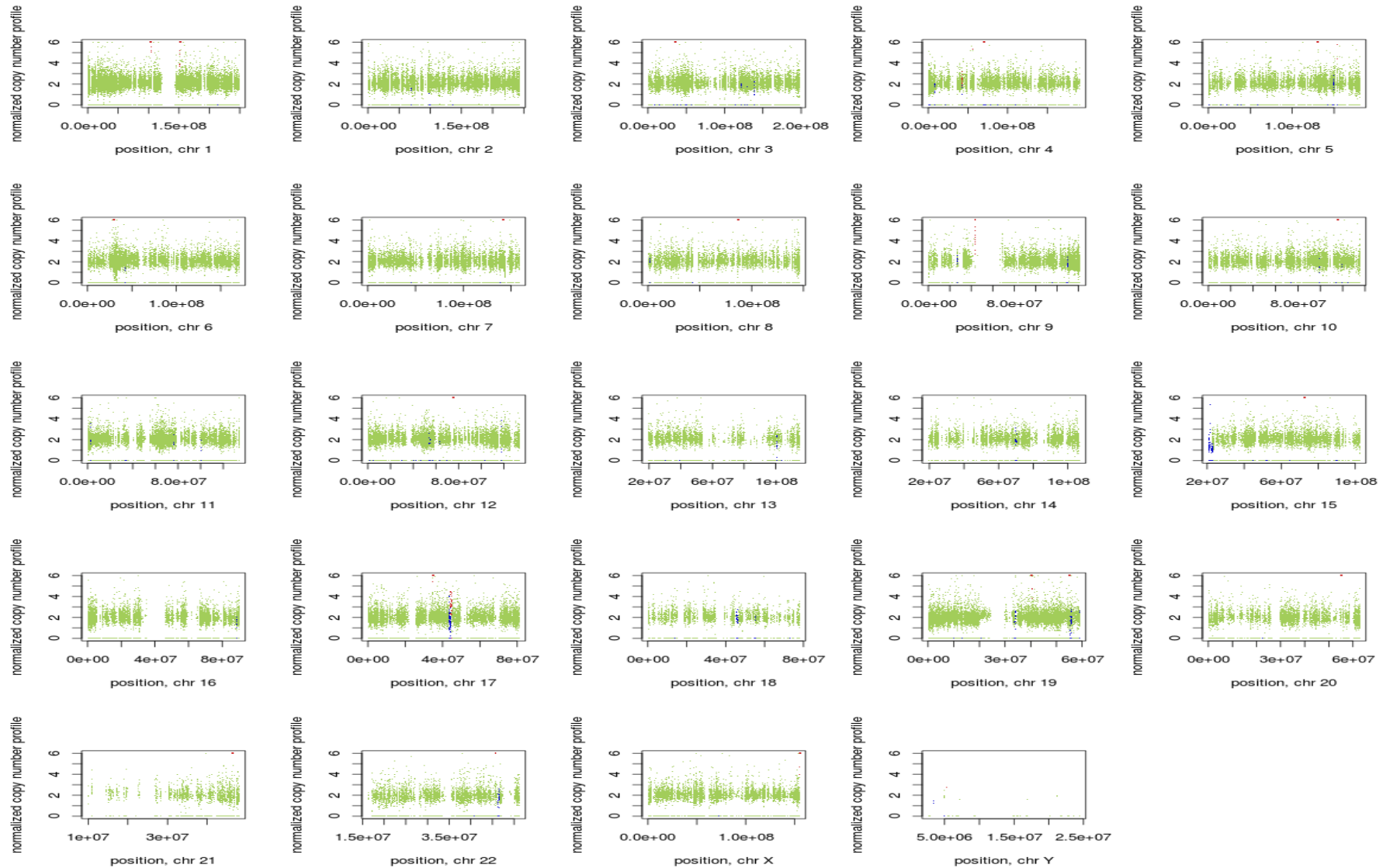


Figure 4.35 Family 5 Individual 1 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all blue chromosome X is expected when compared with a control female). None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Analysis of whole-exome, LRoH and candidate genes

As before, all LRoHs were checked for overlap with known PCD causal genes (Figure 4.36). Following similar filtering criteria as depicted in Figure 4.16, 13 homozygous Φ mutations were identified in list 1 and an additional 314 homozygous Φ variants were identified in list 2 for patient 2. However all mutations in list 1, and all in list 2 were common except one; a missense mutation in the *MNS1* gene (p.M263T, see Figure 4.37). The mutation was absent in dbSNP, ExAC, 1000GP and our internal database, and was present in EVS with a total MAF of 0.0077% (0.0117% in European Americans and none in African Americans). However none was homozygous for the allele. Assuming the individuals participating in EVS follow a Hardy-Weinberg equilibrium, then the probability of observing a homozygote (i.e. $q^2 = 0.000077^2$) for the mutation would be approx. one in two hundred million which could be the explanation of why the mutation was not identified before. The mutation is located within a long autozygous region (~37Mb) and resides in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 664.7 [310]. The missense mutation was predicted to be 'deleterious' by SIFT (score: 0.03) [243], Polyphen-2 (0.932) [244] and Condel (0.743) [257]; and to an extent by (unweighted) FATHMM (-0.79) [242]. A final analysis was carried out on all the mutations outside of the two candidate lists and 5 more missense mutations were identified which were all predicted to be 'tolerated' by one or more of the four mutation prediction tools we used in the analysis; and all were in genes which had no prior connection to any type of cilia. Furthermore, three of these variants were present in a homozygous state in several of our internal controls. The additional CNV analysis yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

This analysis yielded a novel candidate gene for PCD which require follow up via further studies.

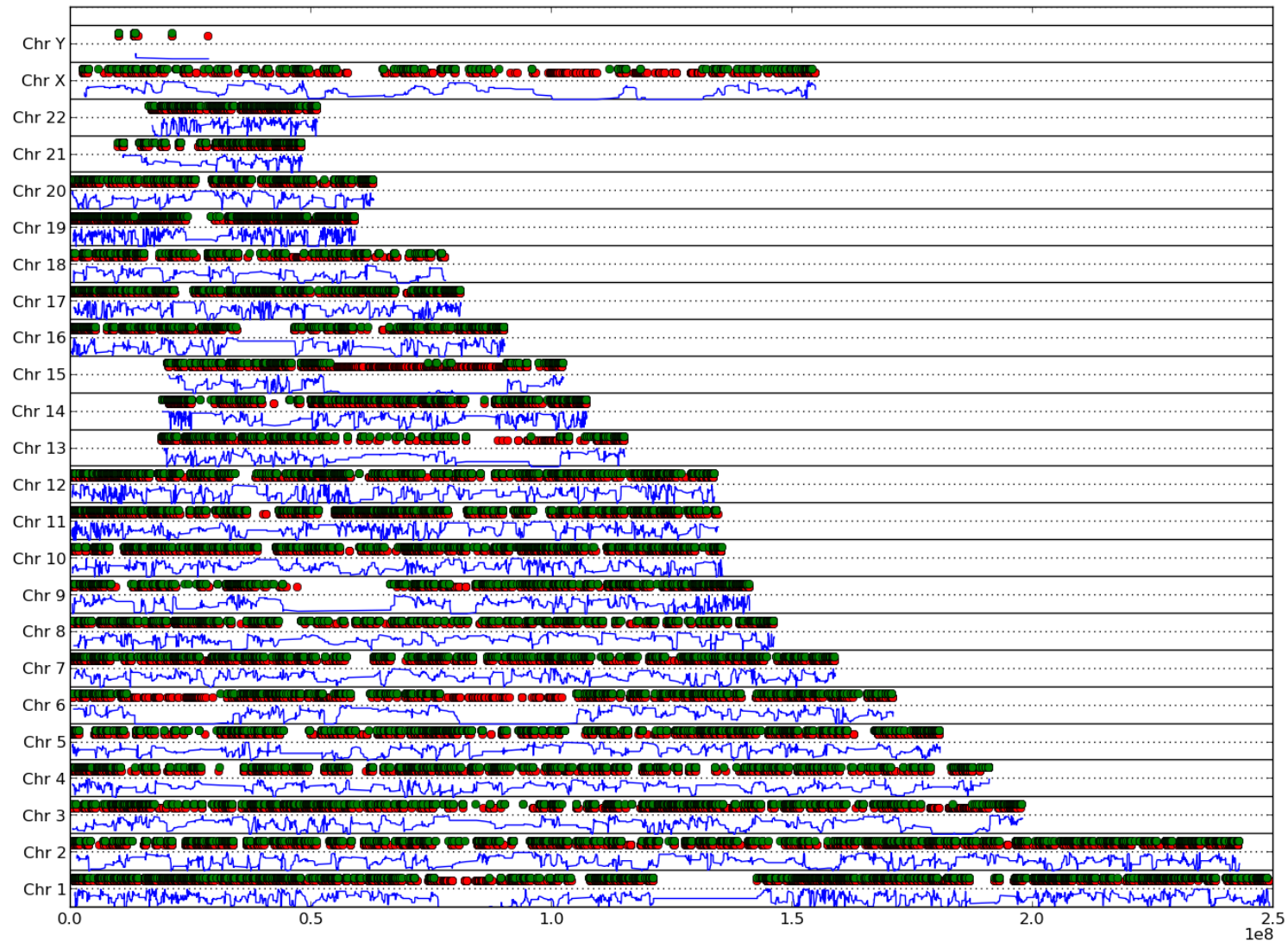


Figure 4.36 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 6 has an autozygous region ~30Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

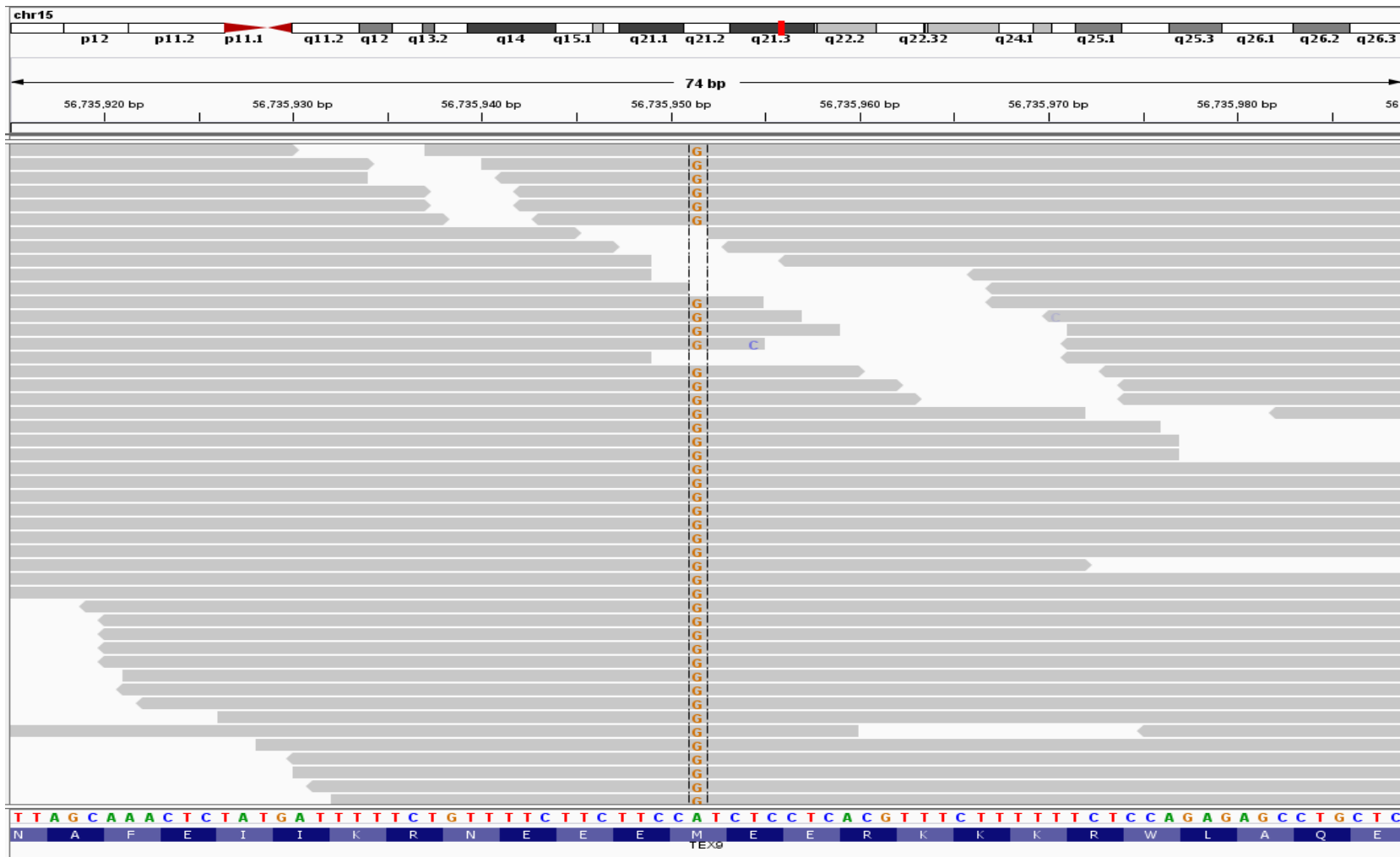


Figure 4.37 Reads mapped to the reference human genome hg19 at the site of the p.M263T mutation. Patient 2 had all 93 reads matching the variant (not all shown). The image was created using IGV [284]. NB: The *MNS1* gene is on the reverse strand and overlaps with the *TEX9* gene (which is on the forward strand).

Modelling mutation effect to protein

Uploading the amino acid sequences stated in section 4.4.9 to the Robetta server yielded 5 different models for both the mutant and the wild type MNS1 protein. Figures 4.38 and 4.39 below depict the first model for both. The affected residue is labelled in both.

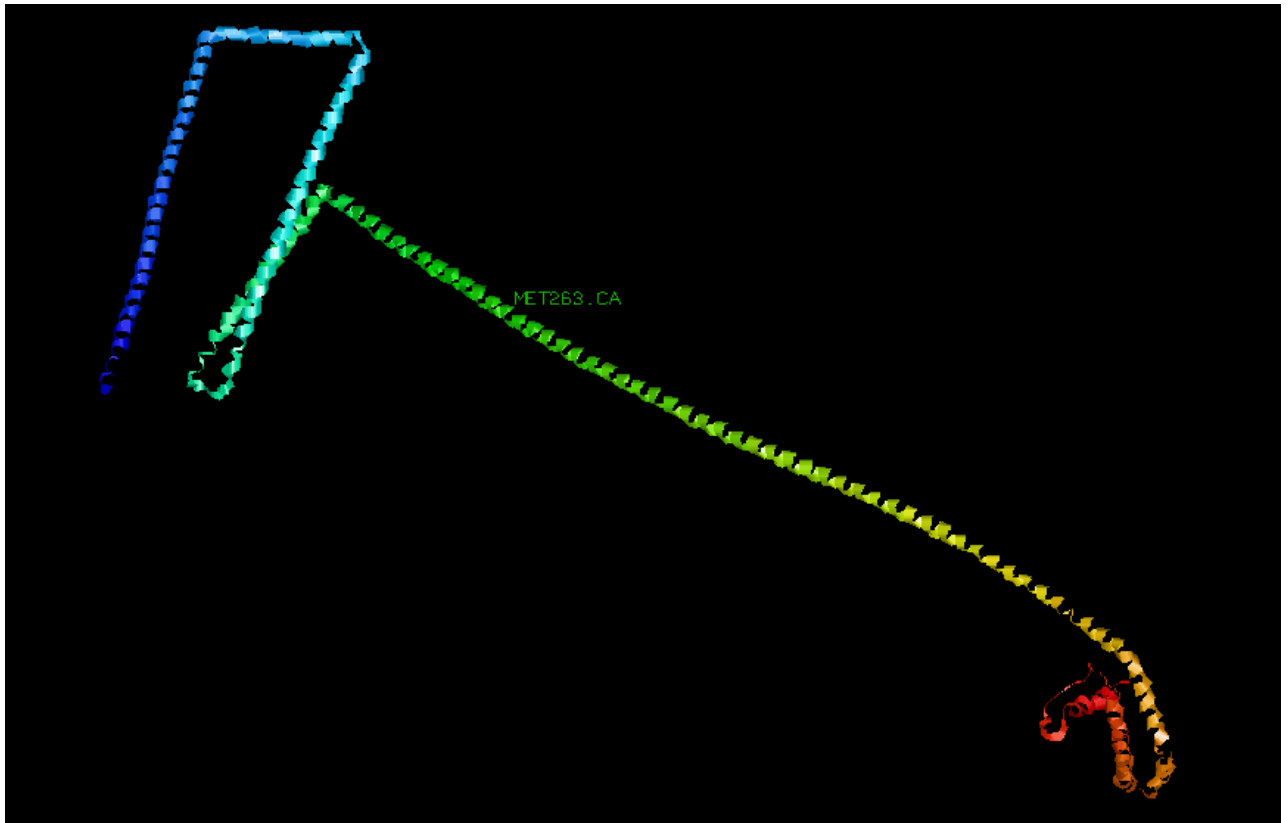


Figure 4.38 Protein structure of wild type MNS1 protein. Only the first (most likely) model is shown and the location of the mutation (p.M263T) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.

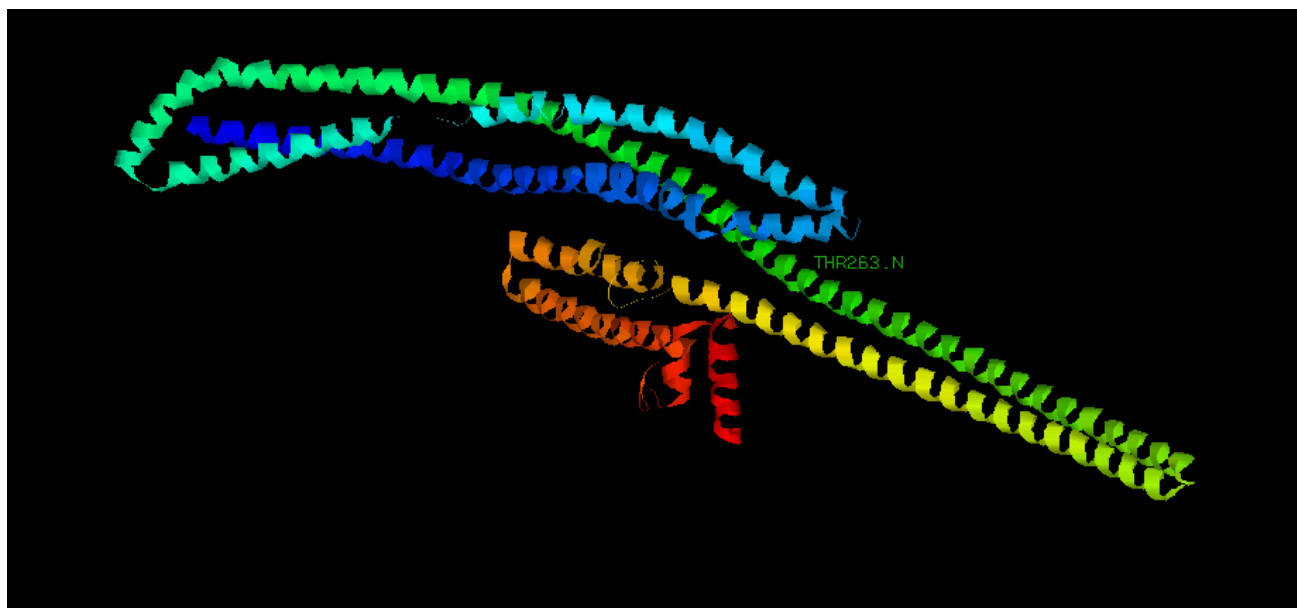


Figure 4.39 Protein structure of mutant MNS1 protein (p.M263T). Only the first model is shown and the location of the mutation (p.M263T) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol. See discussion on section 4.6.2 for details. Comparing the mutant and wild type forms of the MNS1 protein shows that the former has more folds. See discussion on section 4.6.2 for details.

4.5.7. Family 6

This section will present the results for family 6 and describe the findings:

No of whole-exome sequenced participants: 1

No of other family members with DNA available: 3 (i.e. both parents and unaffected sibling)

WES Data statistics

Total captured region was 118,507,605 base pairs (50,620,566 bases on target and 67,887,039 bases near target, the latter being flanking regions within 200bp of exons). Coverage of target (i.e. exons) and flanking regions (e.g. introns, splice sites) was 98.2% and 92.7% respectively. The average sequencing depth on target was 61.49 and the fraction of target covered with at least 20 and 10 reads was 78.5% and 88.2% respectively (and >4 read depth = 94.4%). There were a total of 51,751,389 (high quality) reads with a mapping rate of 99.21%.

Table 4.34 summarises the data produced by the WES platform for individual 1.

Type	Raw data	Clean data
Number of Reads	57402118	54753906
Data Size	5740211800	5475390600
N of fq1	151031	44465
N of fq2	1947749	105397
GC(%) of fq1	44.35~44.37	44.18~44.21
GC(%) of fq2	44.5~44.52	44.32~44.34
Q20(%) of fq1	96.32~96.34	97.19~97.22
Q20(%) of fq2	93.00~93.01	95.85~95.89
Q30(%) of fq1	89.96~90.05	91.07~91.17
Q30(%) of fq2	86.40~86.47	89.22~89.32
Discard Reads related to N	69528	
Discard Reads related to low quality	2368222	
Discard Reads related to Adapter	210462	
Clean data/Raw data	95.39%	

Table 4.34 Family 6 Individual 1's WES data quality summary statistics. Majority of the base calls produced were of the highest quality as presented by the Q20 and Q30 rows. GC content of the reads was also within expected values. Clean to raw data ration was also high, indicating the quality of the data produced from the sequencing platform was high.

Read alignment statistics

Table 4.35 (next page) summarises the quality of alignment of the reads produced by WES platform to the reference genome for individual 1.

Exome Capture Statistics	Value
Initial bases on target	51543125
Initial bases near target	73252334
Initial bases on or near target	124795459
Total effective reads	51751389
Total effective yield(Mb)	5093.94
Average read length(bp)	98.43
Effective sequences on target(Mb)	3169.47
Effective sequences near target(Mb)	1191.82
Effective sequences on or near target(Mb)	4361.29
Number of reads uniquely mapped to target	34639503
Number of reads uniquely mapped to genome	45808512
Fraction of effective bases on target	62.2%
Fraction of uniquely mapped on target	75.6%
Fraction of effective bases on or near target	85.6%
Average sequencing depth on target	61.49

Exome Capture Statistics	Value
Average sequencing depth near target	16.27
Mismatch rate in target region	0.24%
Mismatch rate in all effective sequence	0.23%
Base covered on target	50620566
Coverage of target region	98.2%
Base covered near target	67887039
Coverage of flanking region	92.7%
Fraction of target covered with at least 20x	78.5%
Fraction of target covered with at least 10x	88.2%
Fraction of target covered with at least 4x	94.4%
Fraction of flanking region covered with at least 20x	26.6%
Fraction of flanking region covered with at least 10x	47.5%
Fraction of flanking region covered with at least 4x	73.0%
Mapping rate	99.21%
Duplicate rate	4.73%

Table 4.35 Family 6 Individual 1's alignment quality summary statistics. The high values for the coverage of the target regions, the mapping rate and the fraction of targets covered with at least 4x depth indicate that the reads produced from the sequencing platform were of highest quality. The high values for mapping rates indicate that the read alignment procedure worked exceptionally and not many reads were left unmapped to the reference genome.

Variant calling statistics

Tables 4.36 and 4.37 summarises the SNV and indels detected and called for individual 1 respectively. Figures 4.40 and 4.41 depict the indel and CNV distribution throughout the genome as captured by WES platform.

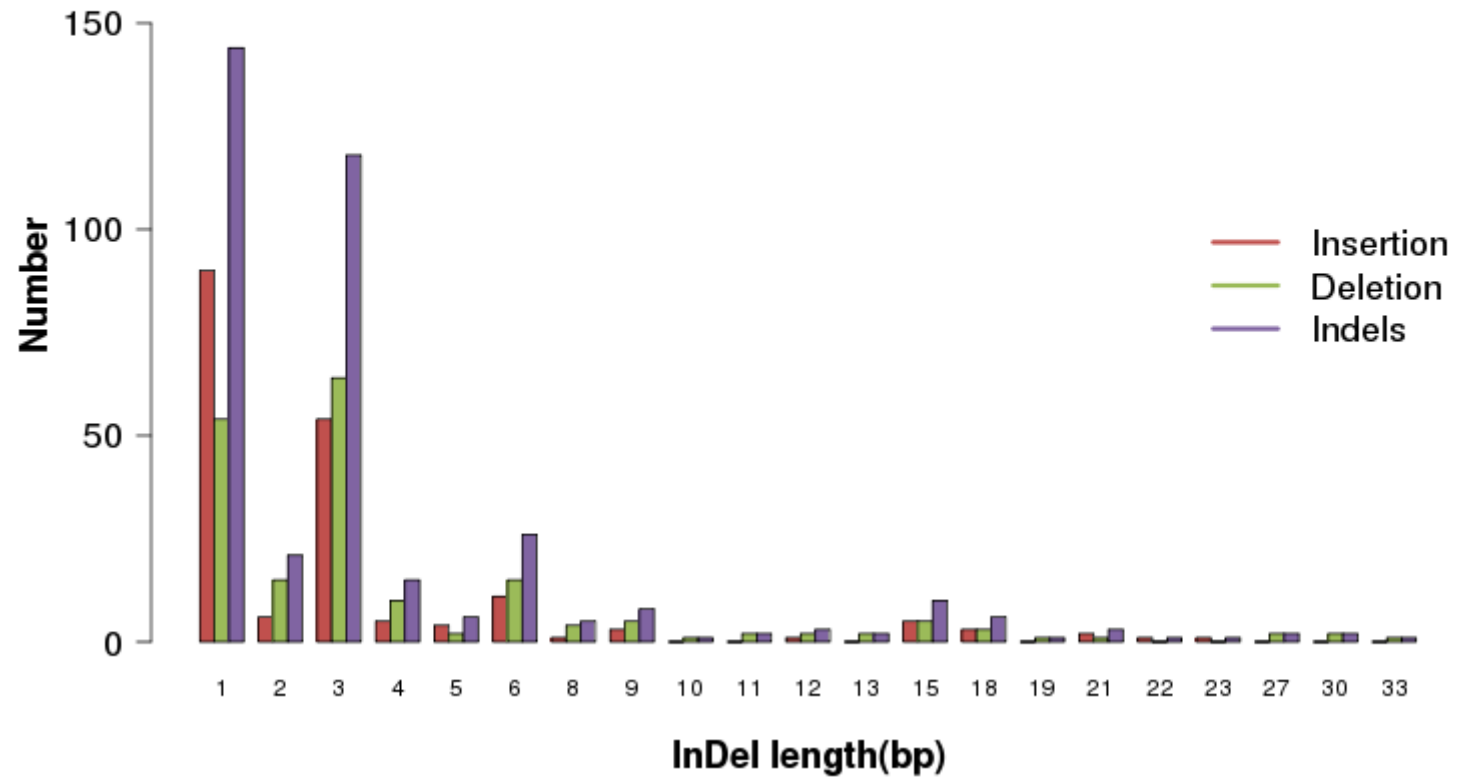
Categories	Value	Categories	Value
Total	78954	Splicing	49
1000genome and dbsnp135	74800	NcRNA	2600
1000genome specific	187	UTR5	853
dbSNP135 specific	2364	UTR5 and UTR3	4
dbSNP rate	97.73%	UTR3	2566
Novel	1603	Intronic	49113
Hom	35396	Upstream	596
Het	43558	Upstream and downstream	49
Synonymous	9710	Downstream	414
Missense	8367	Intergenic	4539
Stopgain	64	SIFT	1010
Stoploss	30	Ti/Tv	2.4240
Exonic	17929	dbSNP Ti/Tv	2.4298
Exonic and splicing	242	Novel Ti/Tv	2.2515

Table 4.36 Family 6 Individual 1's SNP summary statistics. The high values in the amount of variants which are present in dbSNP (row: dbSNP rate) indicate that the variant calling procedure was successful. The Transition to Transversion ratio (row: Ti/Tv) is also consistent with what is expected from WES data (between 2.1 and 2.8).

Categories	Value	Categories	Value
Total	9487	Stopgain	2
1000genome and dbsnp135	4644	Stoploss	0
1000genome specific	978	Exonic	342
dbSNP135 specific	1955	Exonic and splicing	3
dbSNP rate	69.56%	Splicing	34
Novel	1910	NcRNA	324
Hom	4324	UTR5	89
Het	5163	UTR5 and UTR3	0
Frameshift Insertion	88	UTR3	395
Non-frameshift Insertion	77	Intronic	7684
Frameshift Deletion	81	Upstream	77
Non-frameshift Deletion	97	Upstream and downstream	5
Frameshift block substitution	0	Downstream	65
Non-frameshift block substitution	0	Intergenic	469

Table 4.37 Family 6 Individual 1's indel summary statistics. Indels are much harder to map and call than SNPs, thus a reduction in dbSNP rate is expected. See Table 4.36 above for a comparison.

InDel length distribution (CDS)



InDel length distribution (All)

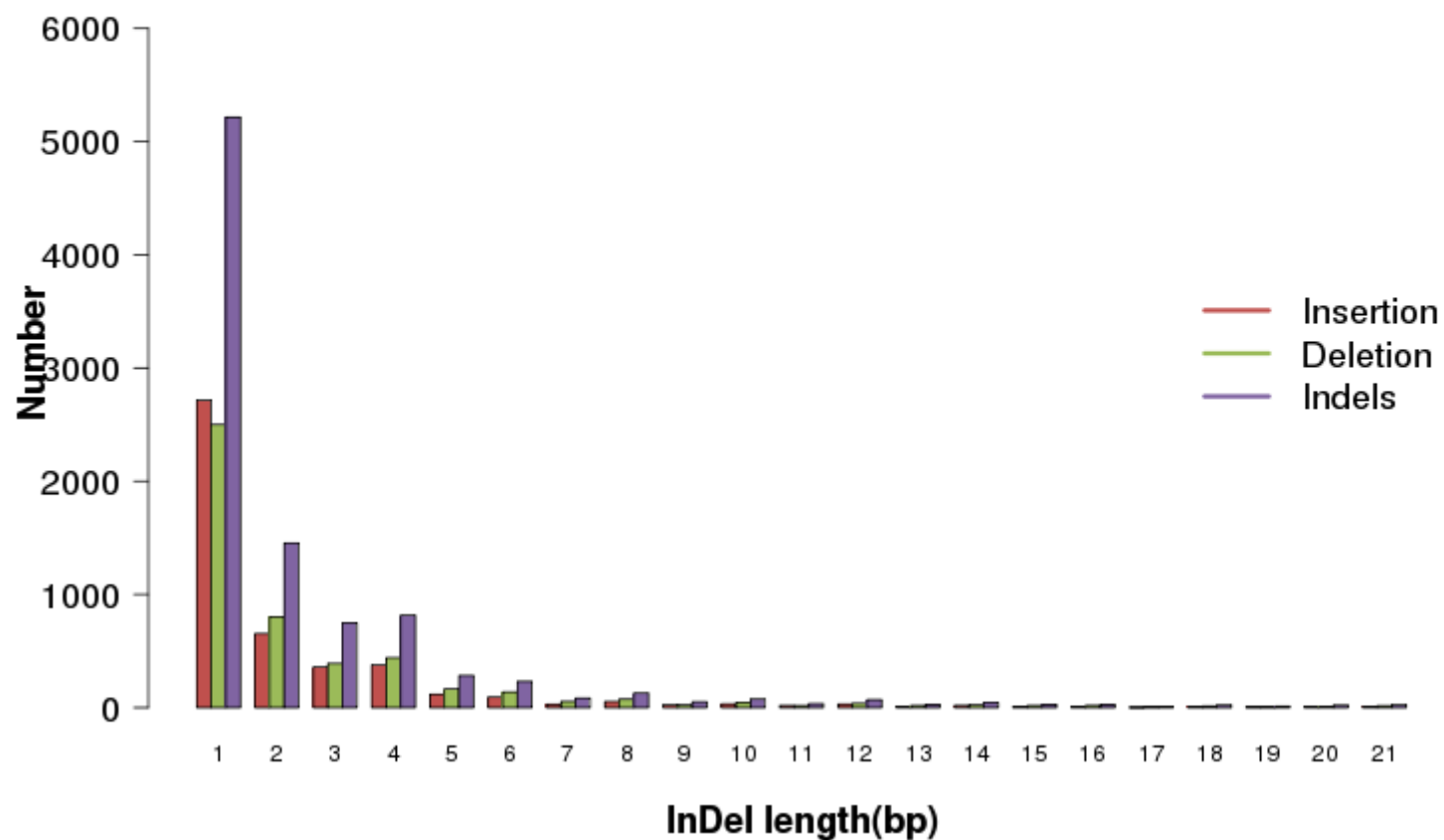


Figure 4.40 Family 6 Individual 1's indel length distributions. The above two figures are complementary (i.e. 'All' and 'CDS'). The top figure is for indels in coding regions (exons) and the latter is for indels in the whole-genome (all that is targeted by the exome-targeting kit). As expected, peaks are consistently observed in indels with size of multiples of 3 (three) as they do not cause frameshifts in the amino acid sequence indicating reliability of variant calling procedure (see indel length distribution within the coding regions, CDS). Peaks at multiples of 3 (three) are not consistently observed in the 'InDel length distribution (All)' figure.

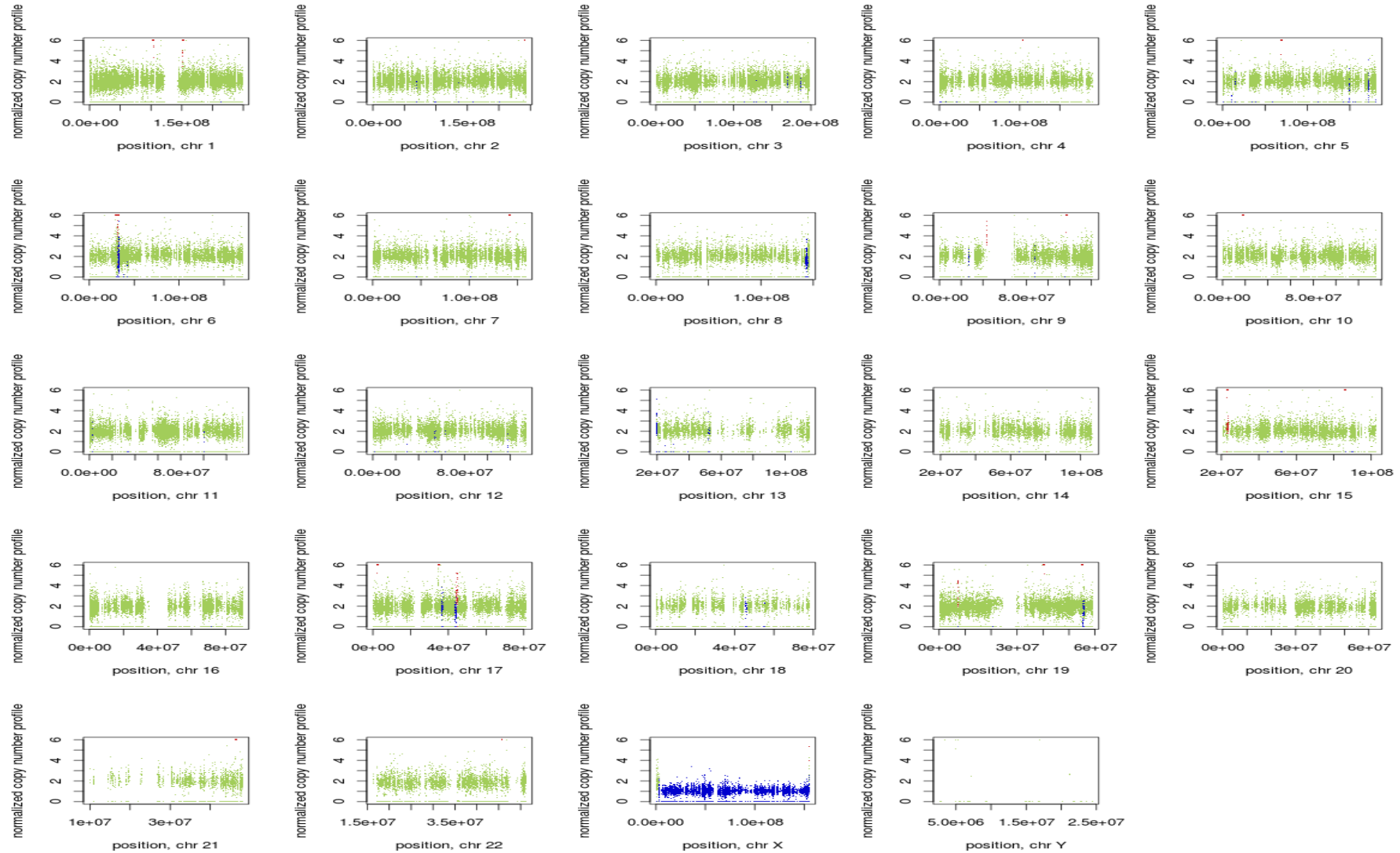


Figure 4.41 Family 6 Individual 1 CNV distribution. Green indicates normal CNV levels, red indicates an increase in CNV levels, and blue indicates a decrease in CNV levels (an all blue chromosome X is expected when compared with a control female). None of the deviations overlapped with an already known PCD causal gene. Image created using makeGraph.R script made available by the Control-FREEC software. NB: Image may have been edited for confidentiality purposes.

Analysis of whole-exome, LRoH and candidate genes

Initially, all LRoH regions identified by Plink [294] and our custom Python script were reviewed in IGV [284] to check whether any known PCD gene (or reported linkage region) resides within the region (Figure 4.42). None of these regions spanned a known human PCD gene or a reported linkage region. Several filters (e.g. MAF, consequence of variant) were applied systematically on all mutations to single out any potentially causal ones. All ‘predicted high impact’ (PHI, hereafter Φ) mutations (i.e. rare stop gains/losses, start losses, splice-site acceptor/donor variants, missense mutations, indels – both non-frameshifting and frameshifting) which were in a homozygous state were analysed separately in the two candidate lists. This yielded only 9 homozygous Φ variants in list 1 (i.e. known genes, see 9.5.3) and a total of 349 Φ variants in list 2 (i.e. all suspected ciliome genes, see 9.5.3). However all the mutations in list 1 were common (i.e. with MAF = >1%) and just 2 of the mutations in list 2 passed this MAF criteria, which were a stop gained (c.925G>T;p.(E309*), see Figure 4.47) in the *CCDC151* gene and the other was a frameshifting insertion in the *ZNF595* gene (p.N201fs). The stop gain was absent in dbSNP, EVS, ExAC, our internal database (which includes his unaffected sibling) and 1000GP databases; whereas the insertion was present (in a homozygous state) in many of our previously whole-exome sequenced controls (of Arabic ancestry) and also had a total MAF of 8.6% in EVS. The filtering process is depicted in Figure 4.43. The stop gain was located within a long LRoH region (~17Mb) on chromosome 19. The mutation falls near the centre of the protein (which is 595 amino acids long) and resides in a highly conserved region represented by a (36-way eutherian mammals) GERP score of 245.1 (also see Table 4.38) [310]. A final analysis was carried out on all the mutations outside of the two candidate lists and no other mutations were identified which passed all the filtering criteria used in Figure 4.43. The additional CNV analysis yielded no apparent gains or losses in the genes contained in the two candidate gene lists.

This analysis yielded a novel candidate gene for PCD which require follow up via further studies.

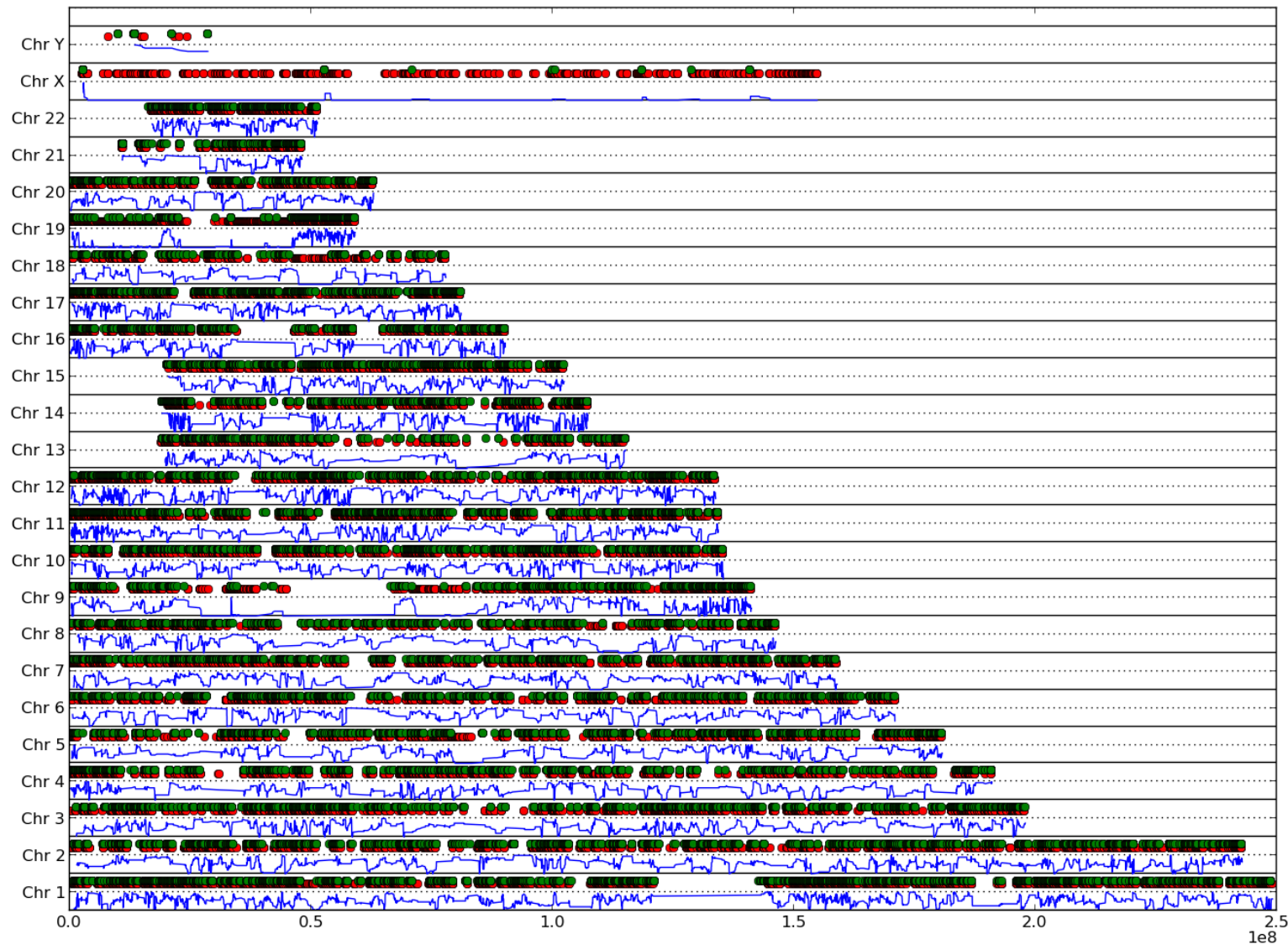


Figure 4.42 Plotting of variant (heterozygosity/homozygosity) status across the genome for individual 1. Green dots: Heterozygous, Red dots: Homozygous. Blue line: likelihood of autozygosity - lower the line, higher the likelihood (e.g. chr 9 has an autozygous region ~10Mb long). The Y axis represents the chromosome number, and the X axis represents the chromosome position. The image was created using AutoZplotter.

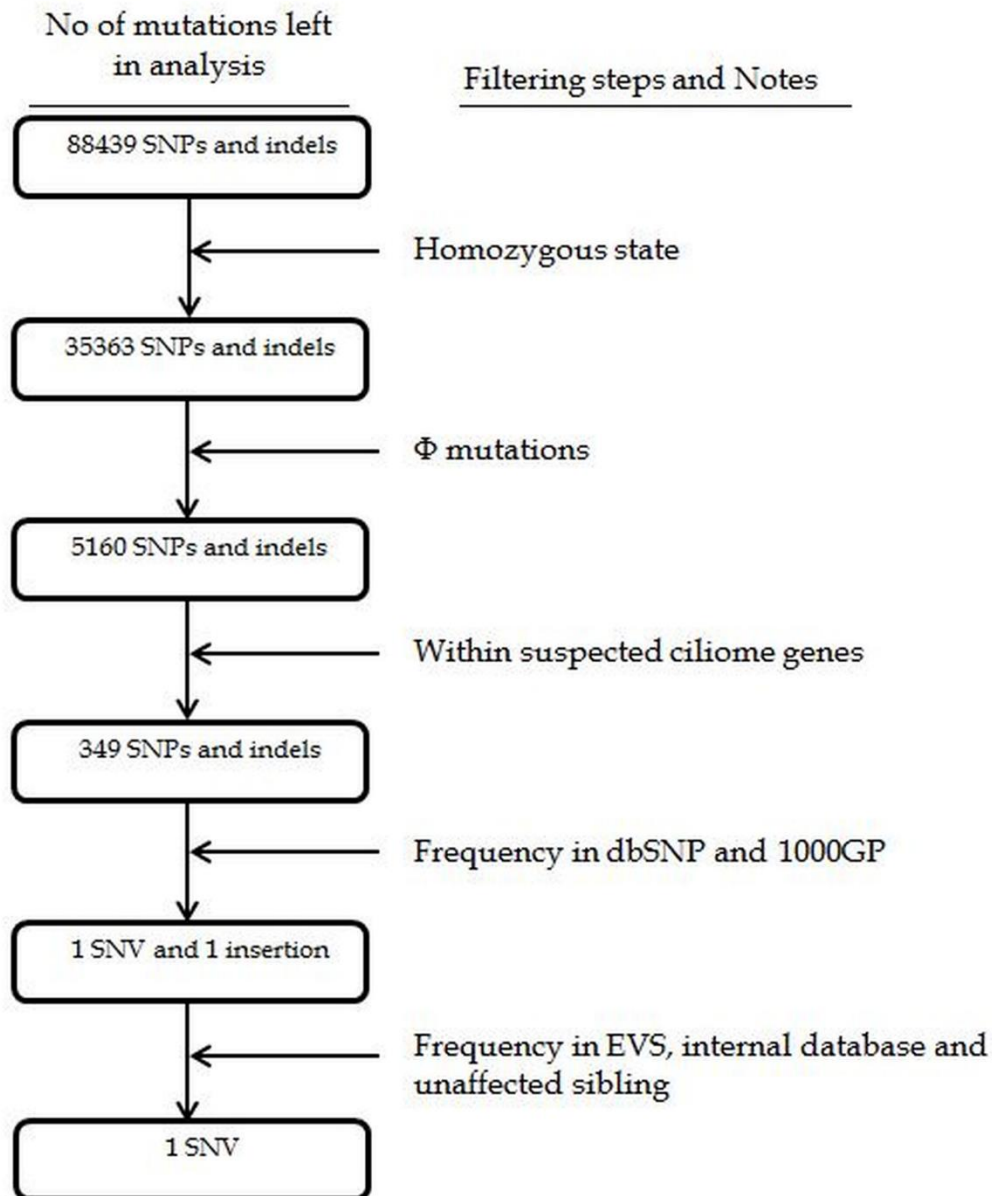


Figure 4.43 Filtering steps applied to all mutations in the exome. After all the filtering steps in the above figure were applied, the total was reduced to a single one in *CCDC151* (GenBank reference sequence: NM_145045.4). Φ mutations: rare stop gains/losses, start losses, splice-site acceptor/donor variants, missense mutations and exonic indels (see section 9.5.3 for details).

Start	Amino acid sequence	End	Entry	Entry Name	Organism
298	ERYISECKKRA E EKKLENERMERK	321	A5D8V7	CC151_HUMAN	Homo sapiens (Human)
294	EHYITDCKKRA E EKKLQTERMERK	317	G3X951	G3X951_MOUSE	Mus musculus (Mouse)
298	ERYISECKKRA E EKKLENERMERK	321	F6XD06	F6XD06_MACMU	Macaca mulatta (Rhesus macaque)
277	ERYISECKKRA E EKKLENERMERK	300	H2QFD9	H2QFD9_PANTR	Pan troglodytes (Chimpanzee)
297	ECYISECKKRA E ERKLENQRMERK	320	F7IJB7	F7IJB7_CALJA	Callithrix jacchus (White-tufted-ear marmoset)
298	ERYVTECKKRA E EKKLENERMERK	321	H0XEE5	H0XEE5_OTOGA	Otolemur garnettii (Small-eared galago)
300	ERFISDCKKRA E EKKLQNERMERK	323	I3NHF3	I3NHF3_SPETR	Spermophilus tridecemlineatus (13-lined ground squirrel)
298	ERYLTECKKRA E EKKLQNERMERK	321	A7MBH5	CC151_BOVIN	Bos taurus (Bovine)
298	ERYITECKKRA E DRKLQNERMERK	321	E2RKK3	E2RKK3_CANFA	Canis familiaris (Dog) (Canis lupus familiaris)
298	ERYITECKKRA E ERKLQNERMERK	321	G1LVJ7	G1LVJ7_AILME	Ailuropoda melanoleuca (Giant panda)
298	ERYITECKKRA E DRKLQNERMERK	321	M3Y1B7	M3Y1B7_MUSPF	Mustela putorius furo (European domestic ferret)
247	ETALTELKAQA E EKKAHAERVERR	270	Q2PEE6	Q2PEE6_CIOIN	Ciona intestinalis (Transparent sea squirt)
265	ERQALDFRKQV E ARKLELERIGRK	288	B4PFS1	B4PFS1_DROYA	Drosophila yakuba (Fruit fly)
204	EFYITDCKKRA E EKKLQTERMERK	227	G3H698	G3H698_CRIGR	Cricetulus griseus (Chinese hamster)
---	E*****E**K****R**RK				

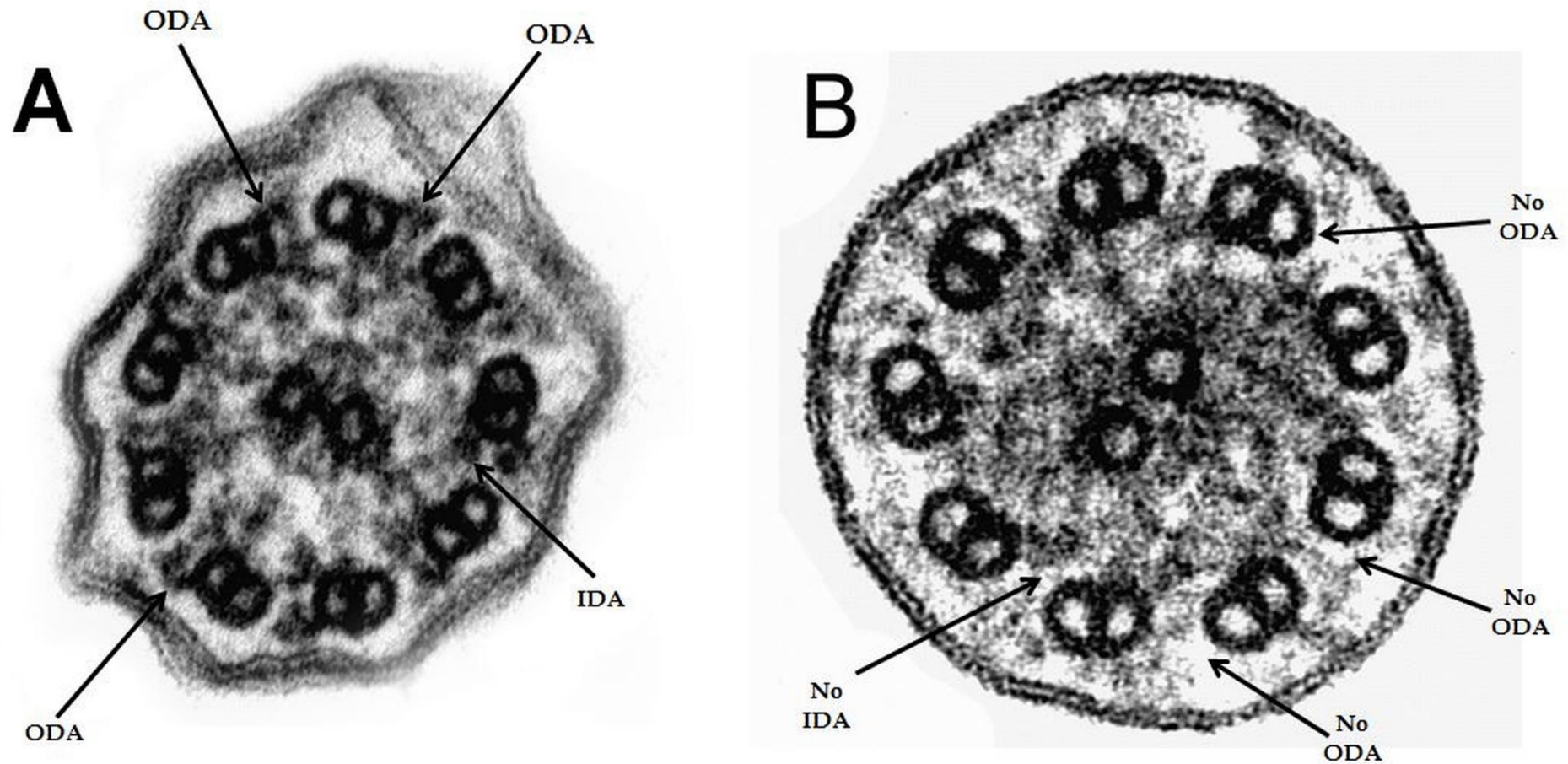
Table 4.38 Local sequence alignment containing the mutated residue from multiple alignment of the *CCDC151* gene in different organisms (relevant species shown). The highlighted Glutamic acid (E) residue is found to be highly conserved across many species where the gene is predicted to be a homologue of the human *CCDC151* gene. The alignment was carried out using the Uniprot website's Blast and Align functions (<http://www.uniprot.org>).

Ultra-structure and Motility of cilia

The EM image of the cross-sections of the mutant and normal respiratory cilia are depicted in Figures 4.44A-B. The EM images clearly show that both the ODA and IDA are missing in the mutant cilia in relation to the ones from a control. Although we cannot provide a video of cilia beating, Dr. Mohammad Mubarak at the King Saud University noted that >80% of the cilia were immotile in the tissue analysed.

Screening for c.925G>T in Saudi Arabian samples

PCR was used to amplify a region 221bp long (harbouring the stop gain) in the proband, the unaffected brother and the parents, which was then subsequently sequenced (using Sanger sequencing method) and digested with *AvrII* enzyme to confirm the variant status (Tables 4.4 and 4.5, Figure 4.45). As expected the parents are heterozygous and the proband is homozygous in accordance with autosomal recessive mode of inheritance of PCD. The unaffected brother is also heterozygous. To deduce how common the c.925G>T:p.(E309*) variant in *CCDC151* (GenBank reference sequence: NM_145045.4) is in the local population, a buccal swab sample from 238 randomly selected individuals of Saudi Arabian ancestry (male and female, living in Riyadh) were collected (see section 4.4.2 for details). The PCR amplicons produced (using primers in Table 4.4) were digested using the *AvrII* enzyme and viewed using 96-well MADGE to check for presence of the p.(E309*) variant. None of the 238 wells showed any digestion (Figures 4.46A-C).



Figures 4.44A-B Cross-sections of respiratory cilia in (A) control and (B) *CCDC151* mutated proband (n=247). The ultrastructural EM images of the cilia confirms the absence of IDA and ODA in the *CCDC151* mutant cilia (74% of 247 cilia scored) – similar to the *ccdc151* morphant in [416]. EM magnification: 250000x

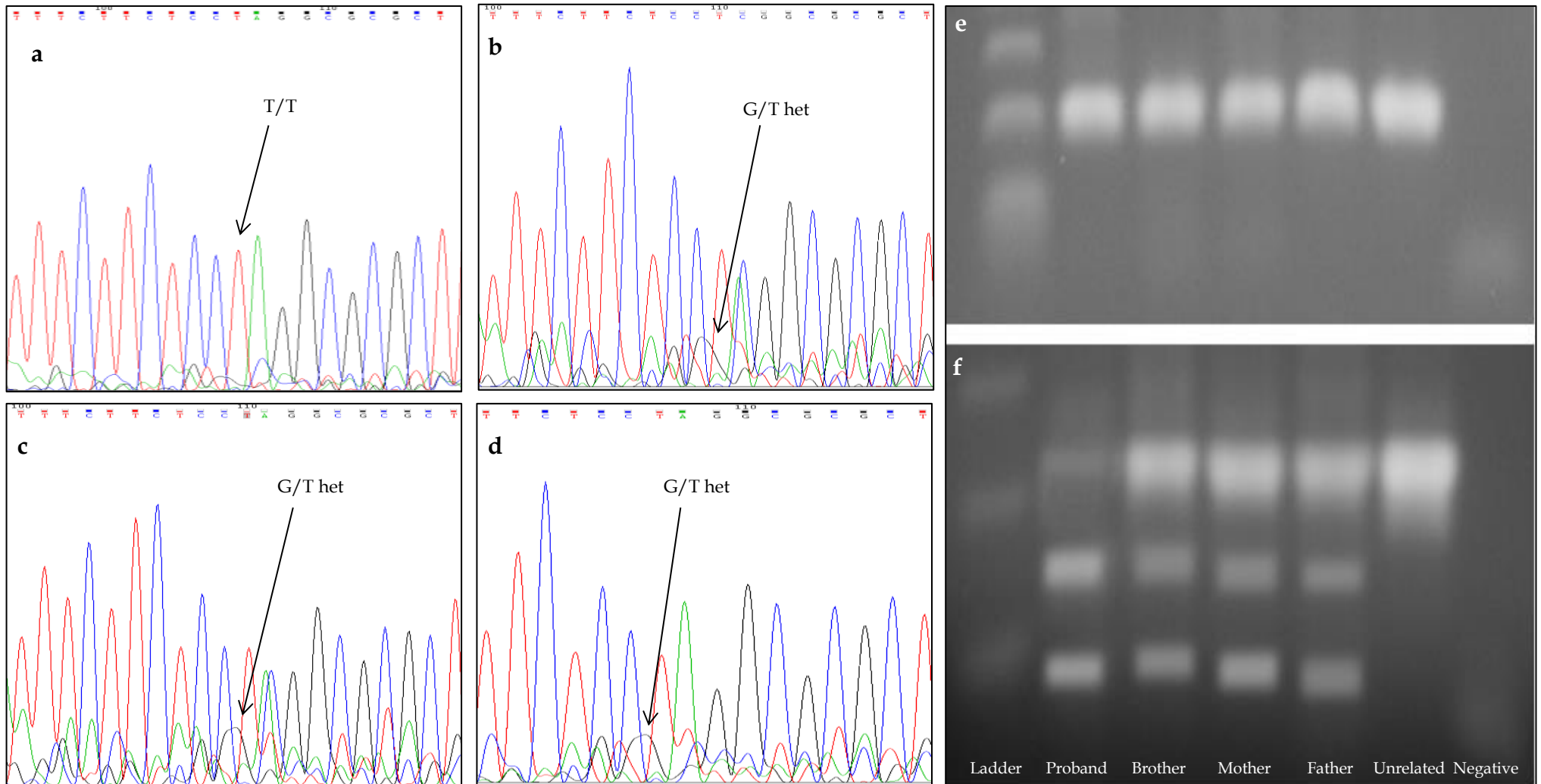
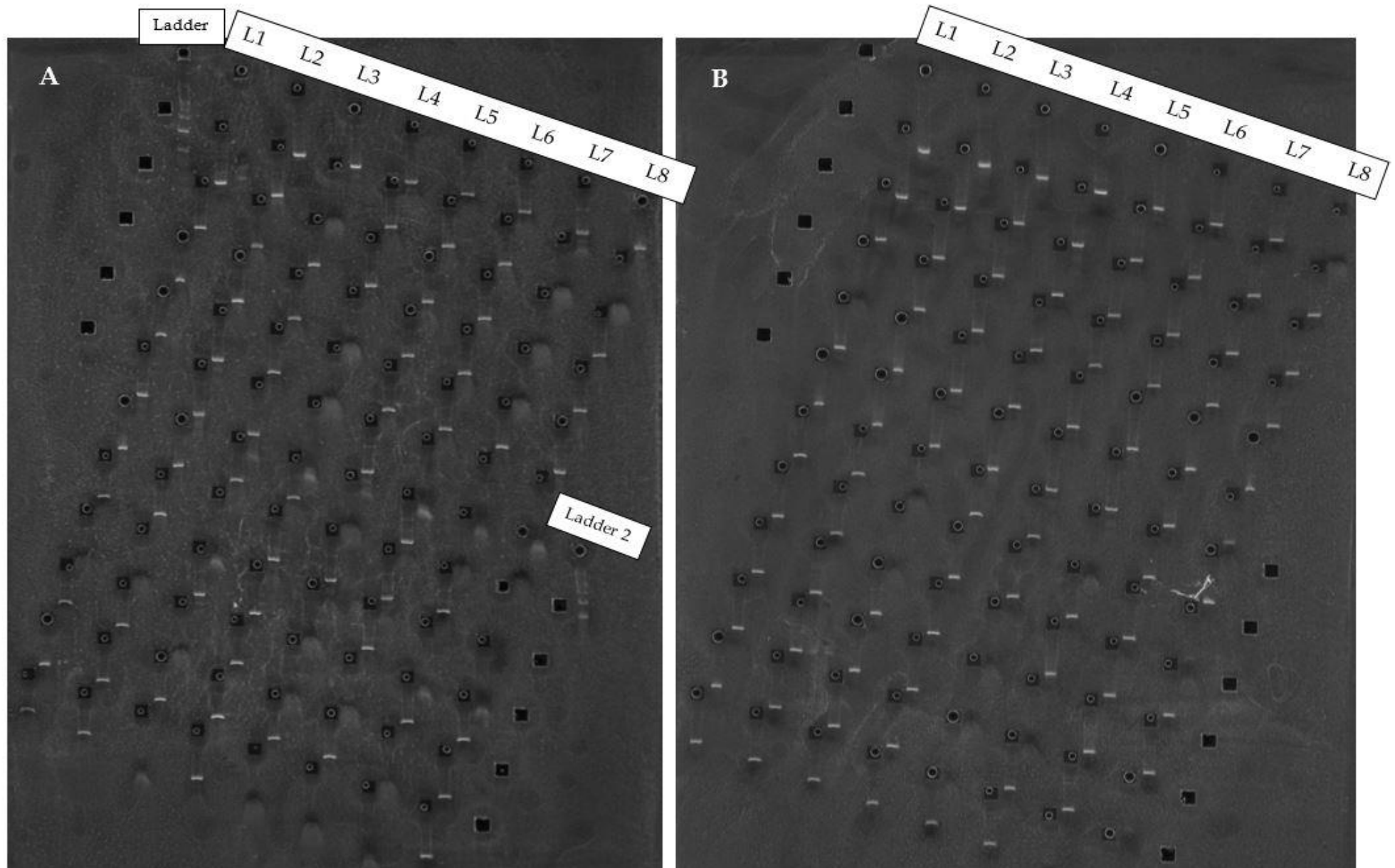
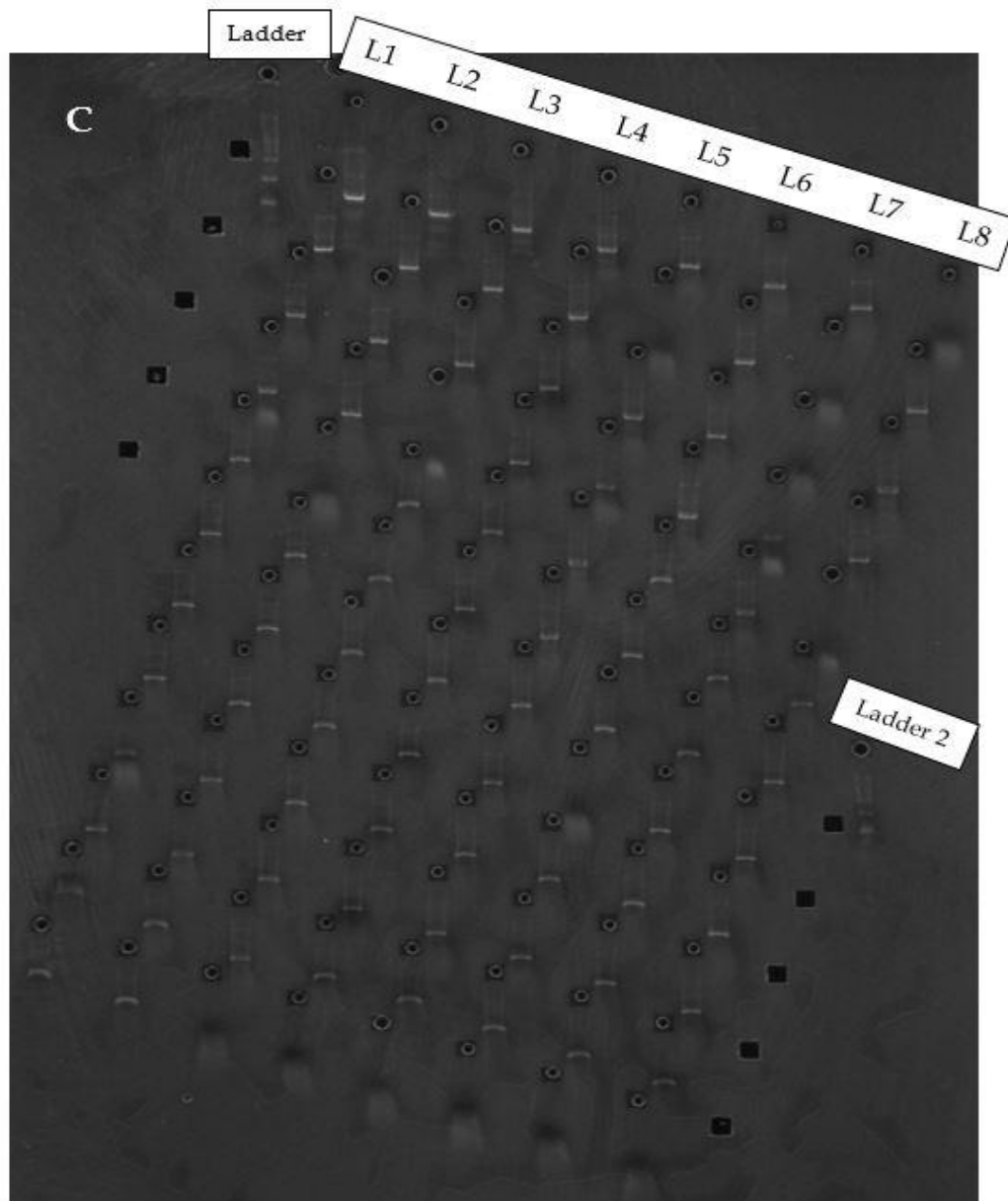


Figure 4.45 Confirmation of variant status in proband and other family members using (a-d) *Sanger sequencing and (e-f) *AvrII* digestion. (a) Proband (b) Unaffected brother (c) Mother (d) Father (e) PCR amplicons before restriction enzyme digestion (f) After digestion. Ladder: 300bp (top), 200bp and 100bp (bottom). DNA Chromatogram images were created using Chromas Lite (v2.1.1). *Peak height imbalances could have been caused by low template DNA, degraded DNA and/or preferential amplification.





Figures 4.46A-C Screening the local Saudi population for the p.E309* variant. 96-well MADGE images reveal that none of the 238 individuals have the causal allele. Ladders last three bands are 100bp (bottom), 200bp and 300bp (top).

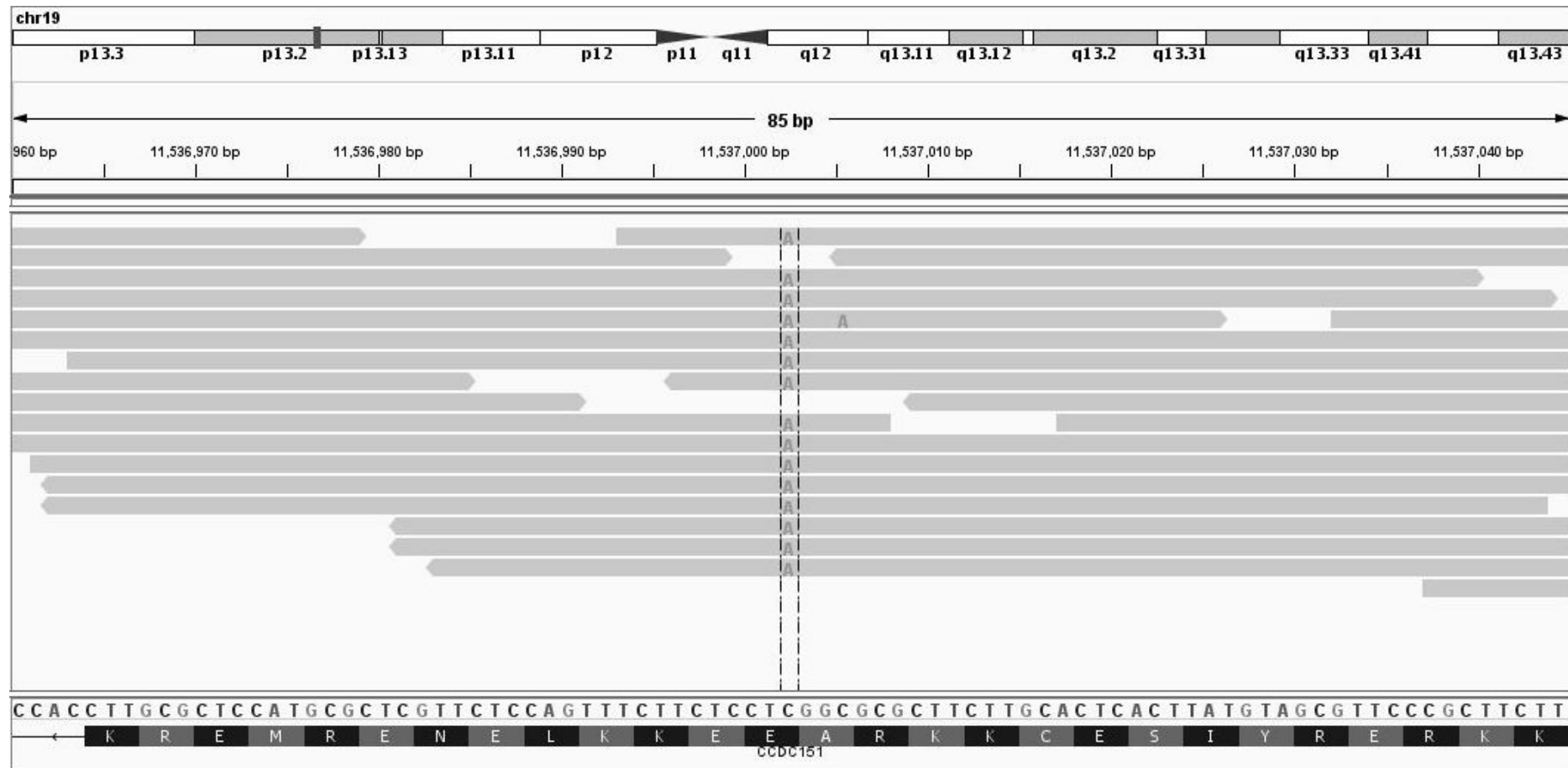


Figure 4.47 Reads mapped to the reference human genome hg19 at loci of the c.924C>A (p.E309*) mutation. All fifteen reads are high quality and there are no wild type alleles. The image was created using IGV. NB: The *CCDC151* gene (GenBank reference sequence: NM_145045.4) is on the reverse strand, therefore the sequence (shown here) must be reversed (e.g. the mutation is G>T causing GAG to become TAG, a premature stop codon). This variant status (i.e. homo/heterozygosity) was confirmed using Sanger sequencing and *AvrII* digestion in the proband, parents and the unaffected brother (Figure 4.45). Nucleotide numbering system uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon (Met) as codon 1.

Modelling mutation effect to protein

Uploading the amino acid sequences stated in section 4.4.9 to the Robetta server yielded 5 different models for both the mutant and the wild type CCDC151 protein. Figures 4.48 and 4.49 below depict the first model for both. The affected residue is labelled in both.

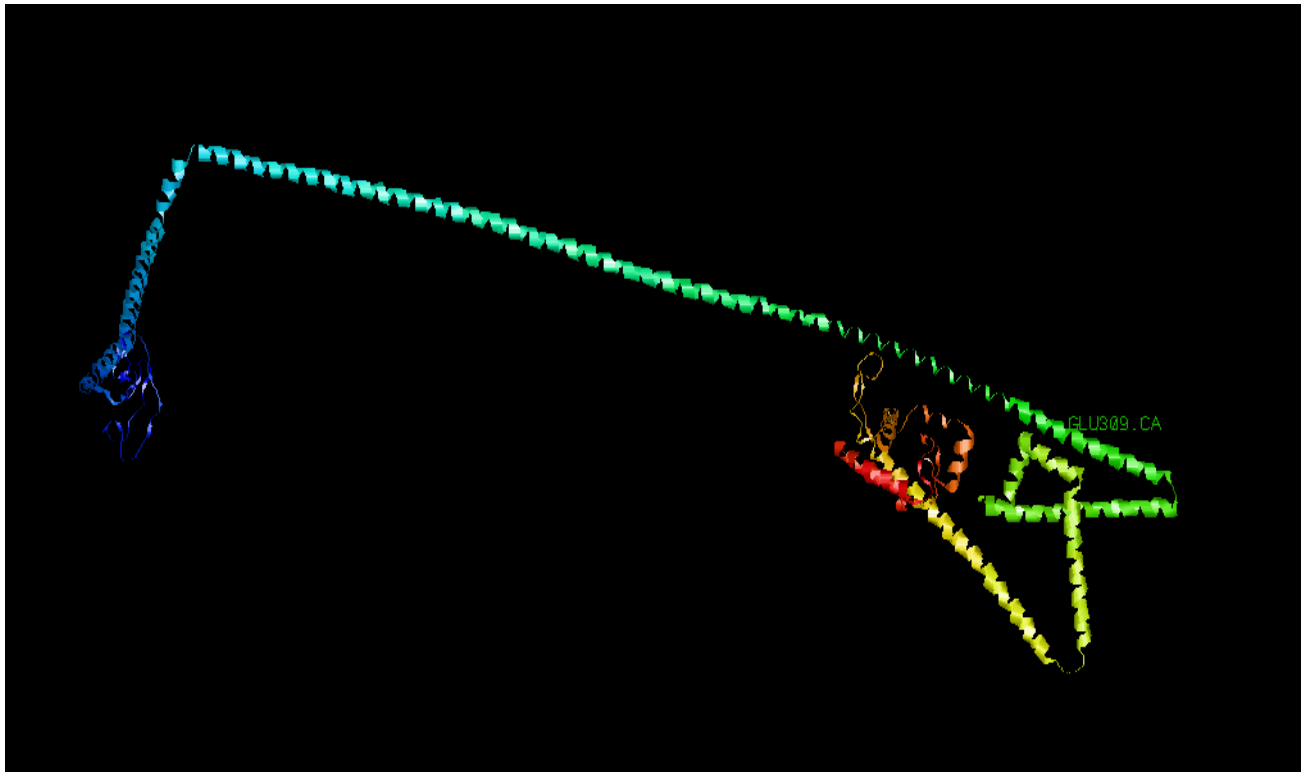


Figure 4.48 Protein structure of wild type CCDC151 protein. Only the first model is shown and the location of the p.E309* mutation is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.

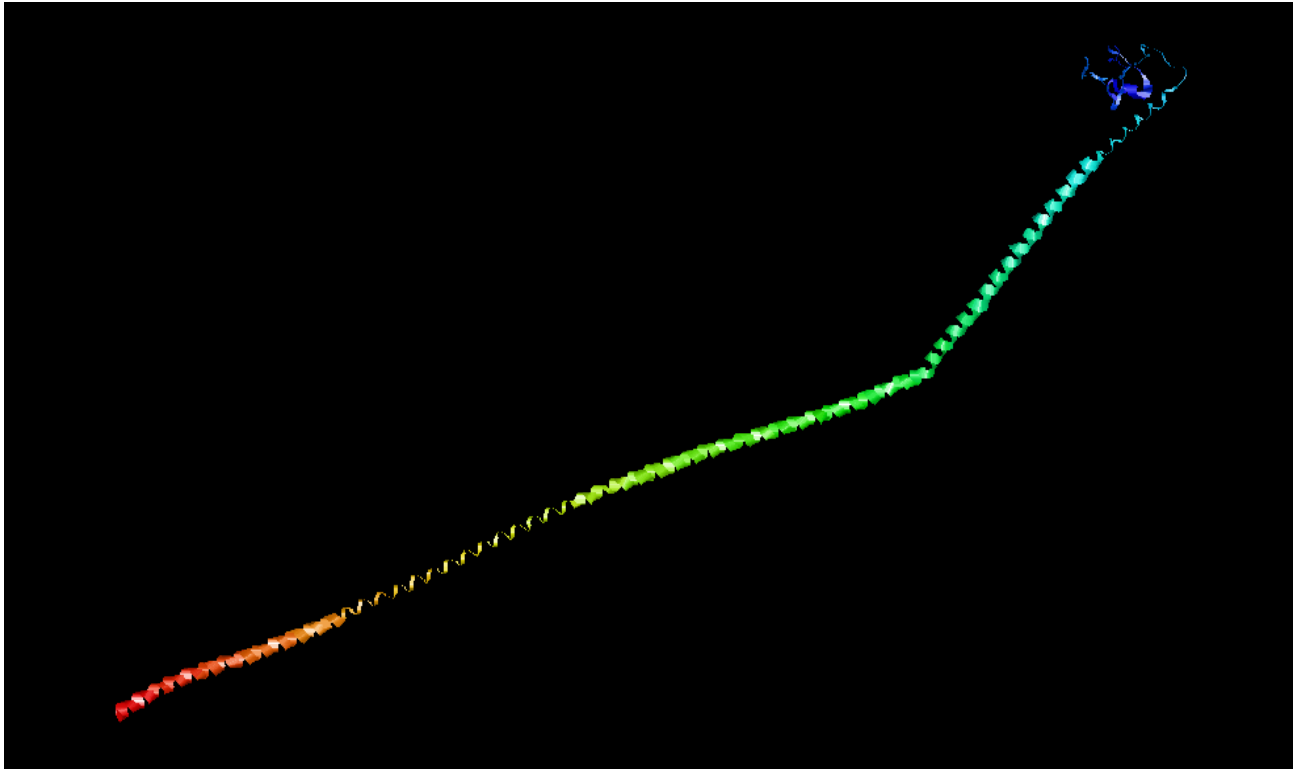


Figure 4.49 Protein structure of wild type CCDC151 protein. Only the first model is shown. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol. The location of the p.E309* could not be labelled as the new protein product is truncated from the stopgain locus (if synthesised, see Figure 4.48). See discussion on section 4.6.2 for details.

4.6. Discussion

In the following section, I discussed the findings in this chapter and also commented on the limitations and caveats of the analyses carried out here. I have included an ‘anecdotes from the literature’ section where I have picked out interesting information from the literature and speculated on how these can relate to the positive and/or null findings in this chapter. I have also made remarks on future work where necessary.

4.6.1. Whole-exome sequencing

Whole-exome sequencing (WES) was the preferred sequencing method in this chapter. Albeit there are natural fluctuations in data quality of the WES carried out, except for individual 1 from family 3, all WES data were of very high quality – reflected by the fact that over 88% of aligned reads had a read depth of 10 or more in

the other 8 individuals. Another reflection of high WES data quality was that the causal variants have been identified in most families, leading to identification of novel genes (i.e. *CCDC151*) or replicating others' findings (i.e. *DNAAF3* variant). In family 1, although a conclusion has not been reached on the PCD causal gene, the same WES data was used in Chapter 6 to carry out a proxy molecular diagnosis for a second disorder (i.e. Papillon-Lèfevre syndrome) indicating the quality of the data for this family was not a limiting factor when looking for the PCD causal variant.

For individual 1 from family 3, the DNA sample was nearly completely degraded (see Figure 10.4 in section 9.5.1 for agarose gel photo after electrophoresis of sample); therefore a lower quality in relation to other samples was expected. Figure 4.33 portraying the CNVs throughout the exome clearly shows the difference between the quality of this sample and the sample of individual 1 from family 4 (the latter was used as the control). However, although the overall read depth was lower, the overall WES data was of sufficient quality for analysis; and a couple of highly likely causal mutations were identified.

As presented in Table 1.2, WES does not pick up variants in non-coding regions, also cannot dissect large CNVs, translocations and structural changes. These can all, especially non-coding variants, cause LoF of certain PCD causal genes hence the causal variant will not be identified when WES is carried out. Other reasons for 'non-detection' may be the less-than-perfect performance of the read alignment, variant calling and variant annotation tools used here. For example, O'Rawe *et al* showed that concordance of multiple variant-calling pipelines (e.g. GATK, SAMtools, SOAP) was low, indicating some variants may be missed depending on which tool is used [417]. Any of the abovementioned reasons can be a contributing factor for the failure to identifying a causal variant (e.g. in the studies carried out on family 2).

The length distribution of indels in the coding regions shows that peaks are present in lengths which are multiples of 3 (e.g. 3, 6, 9). This is expected as these are non-frameshifting indels which have relatively small effects on the primary structure of proteins in relation to frameshifting indels.

4.6.2. PCD analyses

Although many PCD causal genes have been identified in humans (including a novel one in this thesis), a comprehensive PCD gene interactome has not yet been established. Also the molecular mechanisms of how mutations in certain PCD causal genes also cause laterality defects (e.g. situs inversus), certain combinations of ultrastructural defects and/or sterility (or sub-fertility) is far from being completely understood [418]. However, identifying all PCD causal genes will contribute towards understanding the mechanisms (e.g. interactome of PCD causal pathway) behind these various clinical phenotypes and speed up the initiation of improved intervention trials to treat this complex disorder.

Using WES data of 9 individuals from 6 different families, rare homozygous nonsense mutations c.925G>T:p.(E309*) in *CCDC151* (NM_145045.4) and c.406C>T:p.(R136*) in *DNAAF3* (NM_178837.4) were found to be causal of PCD. Other variants such as c.788T>C:p.(M263T) in *MNS1* (NM_018365.2), c.787C>T:p.(R263*) in *DNALI1* (NM_003462.3), c.2200delG:p.(G734fs) in *HEATR2* (NM_017802.3) and c.982G>T:p.(E328*) in *LRRC48* (NM_031294.3) have also been identified which may be causal of PCD but studies in this thesis remained inconclusive due to various reasons (e.g. lack of sufficient DNA samples from other family members; loss of contact with participants). We did not have access to the parents' DNA samples for families 3, 4 and 5 which was the main limitation of the studies carried out in these families; although the causal mutation was successfully identified in family 4 (i.e. c.406C>T:p.(R136*) in *DNAAF3*). Additionally, a causal mutation could not be identified in family 2, indicating mutations in non-coding regions or large deletions/insertions which are not picked up by WES.

The genes and variants identified in these analyses will be discussed in their respective sections below.

DNAAF3

Mitchison *et al* published *DNAAF3* as a PCD causal gene in 2012. The p.R136* mutation was also one of the mutations identified in their studies (alongside p.V255Cfs*12 and p.L108P) [34]. Therefore this is conclusive evidence that this was the causal mutation in the family analysed in this thesis (see section 4.5.5). In their study, mutations in *DNAAF3* caused PCD with immotile cilia that lack both outer and inner dynein arms. They have also carried out functional studies in model organisms such as *Chlamydomonas* and Zebrafish; and based especially on their *Chlamydomonas* results, they concluded that the cytoplasmic assembly of axonemal dynein motors requires at least two steps (see Figure 8 of the Mitchison *et al* paper [34]):

- (i) An earlier step required for dynein heavy chain stability that may involve the folding of globular dynein head domains
- (ii) A later step that generates an assembly competent complex between dynein heavy chains and smaller subunits

As the p.R136* mutation in *DNAAF3* was a previously reported variant, it is also worth following up through population based studies in the Arabian Peninsula as it may turn out to be relatively common variant (e.g. founder mutation) which can then be included in genetic screening tests carried out in the respective populations and/or countries (e.g. Saudi Arabia).

Comparing the predicted protein structures of the wild type and mutant forms of the *DNAAF3* protein reveals the extent of the truncation caused by the p.R136* mutation. It is easy to observe that most of the protein domains have been removed and therefore it is easy to comprehend how such a large structural change can affect the correct functioning of the protein. Also the stop gain falls towards the centre of exon 4 and more than 50bp remain until the penultimate exon-exon junction in the 'mature' transcript, thus it is highly likely that the mutated transcript will be a target for NMD [419]. If it is, then the whole gene will be inactivated (i.e. no gene product

will be synthesised), rendering the mutant type protein structure predictions shown here redundant.

DNALI1

Biological plausibility of *DNALI1* as a novel candidate gene for PCD

Screening of cilia-related literature clearly sets *DNALI1* as a potential PCD causal gene. Many large-scale human cilia proteomics and model organism studies (from *Chlamydomonas reinhardtii* to *C. elegans* to *Mus musculus*) have been carried out within the last 10 years; and *DNALI1* was almost always reported as an essential component of respiratory cilia in these studies [345, 420-426]. Additional to these studies, when Zariwala *et al* identified *ZMYND10* as a PCD causal gene, they also found that mutations in this gene resulted in the absence of the axonemal protein component *DNALI1* (alongside *DNAH5* – another known PCD causal gene) from respiratory cilia [405]. Another study which portrays the importance of *DNALI1* was published by Loges *et al* [396]. They showed that *LRRC50* deficiency disrupts assembly of *DNALI1*-containing IDA complexes (alongside *DNAH5*- and *DNAI2*-containing ODA complexes), resulting in immotile cilia [396]. Finally in another study where *DNAAF3* was reported as a PCD causal gene, *DNALI1* was found to be absent from the cilia of affected individuals (i.e. with *DNAAF3* mutations) and mutations in *DNAAF3* completely blocked the assembly of *DNALI1* [34]. Supplementary to all the above mentioned studies, STRING also predicts *DNALI1* to interact with many already identified human PCD causal genes (e.g. *DNAH5*, *DNAI1*, *DNAI2*, *LRRC50*).

Carrying out a comparison between the predicted protein structures of the wild type and mutant forms of the *DNALI1* protein shows that most of the protein remains the same in the latter. Therefore unless the C terminus is proven to be essential to the correct functioning of the protein, it would be hard to accept that the variant itself can be causal on its own. The truncated form of the protein is highly likely to be synthesised as the stop gain falls towards the very end of exon 5 (penultimate) and less than 50bp remain until the next exon-exon junction in the ‘mature’ transcript, thus it is unlikely that the mutated transcript will be a target for NMD.

EM images of cilia indicate *DNALI1* may have a role in doublet formation

Analysis of the EM images of the cilia (see Figures 4.14A-E) from the two affected sisters reveals lack of subfiber B in the doublets alongside the loss of inner and outer dynein arms, although the 9+2 formation is retained. Since *DNALI1* is an IDA component it is expected to have a role in IDA formation, however as of current, it is not clear how it can have a direct effect on doublet formation. As far we know this is the first ciliary ultrastructural defect with these phenotypes in PCD patients.

MNS1

A study carried by Zhou *et al* showed that *MNS1* (Meiosis-specific Nuclear Structural 1) was 'essential' for motile ciliary functions (also spermiogenesis* and assembly of sperm flagella) in mice as *Mns1*-knockout (*Mns1* $\Delta\Delta$) mice displayed situs inversus (and hydrocephalus) which is consistent with the human PCD phenotypes (e.g. Kartagener's syndrome) [427]. They also showed *Mns1*-deficient tracheal motile cilia lacked some outer dynein arms in the axoneme [427].

Additionally, in *Mns1*-deficient sperm flagella, the 9+2 arrangement of microtubules and outer dense fibres were completely disrupted [427]. Prior to this direct functional analysis, *MNS1* (or homologs) was also identified by 6 other studies (of many types, e.g. human cilia, nematode worm, *C. reinhardtii*) as a potential (motile) ciliome gene [345, 420-422, 424, 428].

Comparison analysis of the mutant and wild type forms of the *MNS1* protein shows that the former has more folds. This can alter the function of the protein especially if an active site resides around these extra folds caused due to the mutation.

LRRC48

The *LRRC48* (Leucine Rich Repeat Containing 48) gene was first identified by Li *et al* in a comparative analysis [421], and then by Pazour *et al* in *Chlamydomonas reinhardtii* as a potential ciliome gene [422]. Recently, when Wirschell *et al* identified *CCDC164* as a human PCD causal gene [389], they also observed that all the analysed

* Final stage of spermatogenesis

individuals with PCD carrying mutations in *CCDC164*, the *LRRC48* protein was either absent or was present at severely reduced levels relative to control axonemes in ciliary axonemes.

Further studies are required to confirm whether *LRRC48* is a human PCD causal gene or not. We could not complete our study as our collaborating clinician lost contact with the family, therefore further DNA could not be received from other family members.

HEATR2

HEAT repeat containing 2 (*HEATR2*) gene (now renamed *DNAAF5*) was first identified by Horani *et al* in 2012 where they identified a single missense mutation (p.L795P) inherited in autosomal recessive fashion in 9 related subjects (of Amish background) affected by PCD [399]. Subsequent ultrastructural and functional analyses revealed absence of both dynein arms, loss of ciliary beating and that the airway epithelial cells isolated from the participants had reduced *HEATR2* protein levels. Additional immunohistochemistry studies in these cells showed that *HEATR2* protein were not localised in the cilia ultrastructure, instead at the cytoplasm suggesting a role in either dynein arm transport or assembly.

CCDC151

CCDC151 gene was first identified as a potential human ciliome gene by Ostrowski *et al* in 2002 [345]. The expressed sequence tag (EST) of *CCDC151* was amongst the 110 proteins (with Accession no: BAB01602) identified by 1-dimensional Polyacrylamide gel electrophoresis (1D-PAGE) analysis of human ciliary axonemes (an additional 104 additional proteins were identified using different methods).

More recently, Jerber *et al* carried out functional analyses of *CCDC151* in *Drosophila*, mice and zebrafish [416]. They showed that *CCDC151* was associated with motile intraflagellar transport (IFT)-dependent cilia in *Drosophila*. In the same analysis, they reported that *Ccdc151* was expressed in tissues with motile cilia in zebrafish, and morpholino-induced depletion of the gene product lead to, similar to human PCD

phenotypes, left-right asymmetry defects [416]. They demonstrated that *Ccdc151* is strongly expressed in (and was restricted to) motile ciliated tissues, where it is required for dynein arm assembly and for the transport of the docking complex *Ccdc114* (homolog of a known human PCD causal gene) [416]. It was also required for proper motile function of cilia in the Kupffer's vesicle (a ciliated organ in the *zebrafish* embryo that initiates left-right development of the brain, heart and gut [429]) and in the pronephros (an excretory organ in vertebrates) by controlling dynein arm assembly, showing that *Ccdc151* is important in the control of IFT-dependent dynein arm assembly in many animals [416]. Furthermore, knockdown of *Ccdc151* in IMCD3* mouse cells resulted in a deregulated ciliary length [416]. A similar analysis was carried out by Dean *et al* in *Chlamydomonas* ODA10 gene which is the homolog of the mouse *Ccdc151* gene (very similar to the human *CCDC151* gene, see Table 4.38) and was found to play an important role in the outer dynein arm assembly [430].

Consistent with the abovementioned zebrafish study [416], the cilia structure of the *CCDC151* mutant cilia from the PCD patient is strikingly similar to the *ccdc151* morphant cilia (see Figure 5F of Jerber *et al* [416]) where the axonemes do not assemble a full complement of ODA and IDA (Figures 4.41A-B). These observations support the essential role *CCDC151* plays for targeting of dynein arms and the dynein arm-docking complex to the axoneme [416].

More recently, another group has also reported *CCDC151* as a human PCD causal gene and identified loss-of-function mutations in five affected individuals from three different families [431]. The affected individuals' cilia showed a complete loss of ODAs and severely impaired ciliary beating. They have also carried out functional studies in mice and zebrafish and found similar results to the Jerber *et al* paper presented above.

* Inner medullary collecting duct

It is important to point out that their phenotype is slightly different from ours that they did not observe the loss of IDAs. It would require further analyses and/or collaborations to find out why this was the case.

Comparing the predicted protein structures of the wild type and mutant forms of the CCDC151 protein reveals the extent of the truncation caused by the p.E309* mutation. It is easy to observe that most of the protein is affected and therefore it is easy to comprehend how such a large structural change can affect the correct functioning of the protein. However the mutant form of the transcript is likely to be a target for NMD as the stop gain falls towards the end of exon 7 which is closer to the 5' end than the penultimate exon, thus the gene will be completely inactivated.

4.6.3. Anecdotes from literature

In 1994, Narayan *et al* published an 'unusual' paper* – as they put it [394]. They published their findings from a family where a PCD affected mother had 5 affected children from three men of different descents (see Figure 1 in Narayan *et al* paper for complete family pedigree). They concluded that PCD may be inherited in an autosomal dominant or mitochondrial fashion. Looking at the data only in an unbiased fashion indicates the latter mode of inheritance is the most likely†. Since they did not carry out sequencing, the causal gene and variant was not identified.

A second paper indicating autosomal dominant mode of inheritance of PCD was published by Gonzalez *et al* [432]. These findings still remain as an enigma till this day, however these findings cannot relate to our findings as none of the parents were affected indicating the absence of human PCD causal autosomal dominant or mitochondrial mutations.

* The title of the paper was: Unusual inheritance of primary ciliary dyskinesia (Kartagener's syndrome)

† Probability of autosomal dominant mode of inheritance = $\frac{1}{2}^5 = 1/32$

4.7. Conclusions

In this study we identified several novel PCD candidate genes. They were *DNALI1*, *MNS1*, *LRRC48* and *CCDC151*. *CCDC151* has later been published as a PCD causal gene and the finding was also replicated by another group [431].

It is also possible that PCD may sometimes* be caused in an additive and/or bigenic (or even polygenic) fashion, where different mutations contribute towards the PCD phenotypes. Findings similar to the latter may shift the autosomal recessive paradigm to more refined PCD inheritance models.

In conclusion, the PCD analyses carried out in this thesis yielded a known variant in a known gene (i.e. p.R136* in *DNAAF3*), a novel variant in a known gene (i.e. p.G734fs in *HEATR2*) and novel variants in novel genes (i.e. p.E309* in *CCDC151*, p.M263T in *MNS1*, p.R263* in *DNALI1* and p.E328* in *LRRC48*). Also one family study did not result in a potentially causal variant. Additional conclusions for each variant and gene can be found below.

DNAAF3

In this study, one of the variants we observed in our subjects (p.R136* in *DNAAF3*) was previously identified in PCD patients providing further evidence of its causal role [34] and also raising the possibility of this variant being a founder mutation in the Arabian peninsula (e.g. a relatively common variant which may be present in seemingly unrelated families, which initially occurred in small inbred populations).

DNALI1

Although a bulk of evidence represented *DNALI1* as a clear candidate for PCD, its homozygous knockout effects were unclear as it was not observed in humans before. Here we reported a nonsense mutation (p.R263*) in *DNALI1* in a patient who has been clinically diagnosed with PCD. As expected the parents are heterozygous and

* Although these types of hypotheses are not preferred due to Occam's razor – law of parsimony

the proband is homozygous for the mutation. However her affected sister is heterozygous which is not in accordance with the conventional autosomal recessive mode of inheritance of PCD (Figure 4.15). This indicates that *DNALI1* is haplosufficient as the parents are not affected.

Bigenic inheritance may prove to be the answer with the combination of another mutation which affects mucus concentration or cilia function, a LoF mutation in *DNALI1* can be PCD causal even with a single copy inactivated. Another possibility is mosaicism. Although the individual inherited a single copy of the LoF mutation in *DNALI1*, the precursor lung cells may have been mutated during development causing sufficient number of cilia to be dysfunctional and therefore cause PCD in the heterozygous sibling. However this hypothesis could not be tested due to technical issues (e.g. lack of biopsy material and tools).

Future studies can conclusively add *DNALI1* to the number of already identified PCD causal genes (with many of them being dynein related genes e.g. *DNAH5*, *DNAL1*, *DNAI1*, *DNAAF1*) and can confirm the potentially ‘essential’ role *DNALI1* plays in human respiratory ciliary function. However, at current, it is not entirely clear how a mutation in *DNALI1* can affect the formation of doublets in cilia which deserves further follow up studies.

MNS1

Although over 3 years has passed since *MNS1* was first implicated as a potential PCD causal gene by animal knockout studies [427], a *MNS1* ‘knockout’ has not been observed in humans. To my knowledge, this is the first time a variant in *MNS1* has been associated with PCD in humans. The variant identified is highly conserved and is predicted to be deleterious by many variant effect prediction algorithms. However these are not sufficient evidence for causality, therefore further studies are required especially on consanguineous populations where a homozygous LoF variant is most likely to be observed.

Unfortunately we could not take this study further due to communication issues between the family and the clinician; and therefore no further DNA could be obtained from the family.

LRRC48 and HEATR2

A homozygous stop gain was identified in *LRRC48* in a PCD patient. However the mutation coincided with another predicted high impact mutation in a homozygous state in *HEATR2* – a known PCD causal gene. Therefore it is more likely that the latter is the causal one. However since there is no conclusive evidence, *LRRC48* cannot be ruled out just yet. Additional studies are therefore required.

CCDC151

Although animal models have strongly linked *CCDC151* to cilia [416], complete human inactivations of *CCDC151* were not observed (until now) to associate these findings to the human model. Here we reported a homozygous nonsense mutation (c.925G>T;p.(E309*)) in *CCDC151* in a patient who has been clinically diagnosed with PCD. The variant was screened in 238 unrelated individuals and it was found to be absent in all of them indicating that the mutation is not a founder mutation and may have occurred relatively recently and/or is tribal specific.

Nevertheless, *CCDC151* adds to the already identified 25 genes (with six of them being coiled-coil domain containing genes: *CCDC39*, *CCDC40*, *CCDC65*, *CCDC103*, *CCDC114* and *CCDC164*) in which mutations are known to cause PCD and indicates the important role *CCDC151* plays in human respiratory ciliary function. Our findings also show that given prior knowledge from an animal model, even a single whole-exome sequence (with high read depth) can be adequate when pinpointing a novel causal gene.

CHAPTER 5. MUTATION IN *ADAT3* CAUSES AUTOSOMAL RECESSIVE INTELLECTUAL DISABILITY

In this chapter I have introduced Autosomal recessive Intellectual disability (ARID) as a disease, focusing on what is known about ARID at current, and also presented the findings of this thesis.

5.1. Introduction

Just as in all areas of genetics, next generation sequencing (discussed in section 1.6.2) has facilitated our understanding of the genetic basis of intellectual disability (ID)* greatly with a torrent of genes and variants identified to be causal in the last decade (according to ‘PubMed results by year’, using ‘intellectual disability’ and ‘genetics’ as key words). ID has been defined as:

“A significantly reduced ability to understand new or complex information and to learn and apply new skills (impaired intelligence). This results in a reduced ability to cope independently (impaired social functioning), and begins before adulthood, with a lasting effect on development.”

by the World Health Organization [433]. Diagnosis of ID is based upon three criteria:

(i) sub-average intellectual functioning (ii) limitations in adaptive behavior (e.g. when communicating, self-care, academic skills, leisure) and (iii) symptoms before

* The evolution of the terms used to define ‘intellectual disability’ (ID) is interesting, with ‘idiot’ and ‘mentally retarded’ being widely used in the early 20th century scientific literature – the latter term is still used sometimes but the former is abandoned for the obvious reasons. This implies that ID may also be replaced in the future

18 years of age [434]. However the use of intelligence quotient (IQ) tests are commonly accepted and used to diagnose whether an individual suffers from ID or not, with IQs below 70 used as a threshold [202]. Individuals with IQs between 50 and 69 are diagnosed with 'mild' ID, with IQs down to 35 and 20 considered 'moderate' and 'severe' respectively. Anything below 20 is diagnosed as 'profound' ID [434, 435]. The prevalence of severe and profound ID patients are thought to be below 0.5% but ID as a whole is thought to affect up to 3% of the total Western population, with the economic burden estimated to be tens of billions of dollars in the US only* [435-437]. The influence of environmental factors (e.g. trauma, teratogens) on ID is indisputable but a considerable proportion is also caused by genetic mutations and/or chromosomal abnormalities. However there is still a large portion of ID cases with the cause waiting to be identified (whether due to genetic or environmental factors). For example in 2006, the underlying cause of ID remained unknown in up to eighty percent of the cases [438]. Today it is known that over fourteen percent of cases are due to CNVs (usually over 400 kilobases long) [202, 439]. Also whole chromosomal copy number variations and/or abnormalities account for a considerable portion of ID cases. A well-known example is the case of Down syndrome (trisomy of chromosome 21) which is the most common single cause of ID with an estimated prevalence of between 1 in 750 to one in a thousand [435, 440]. Since the focus of this thesis is 'inherited' ID cases which follow a Mendelian pattern of inheritance (e.g. single-gene causes) and occur in consanguineous individuals (for reasons discussed in 1.2.3 and 1.3), ID cases which are caused by *de novo* means will not be discussed any further. For a comprehensive review on the effects of chromosomal copy number variations on ID, see review paper by Morrow [441].

Fragile Mental Retardation 1 (*FMR1*) gene was the first of many ID causal genes (coined the name Fragile-X syndrome, FXS); and is the most common single gene aetiology of inherited ID [202]. FXS results from a CGG-repeat expansion† that triggers hypermethylation and therefore silencing of the *FMR1* gene – with an

* With average lifetime cost for per person being just over a million dollars

† ~200 repeats is thought to be the threshold for disease

incidence of 1 in five thousand males which accounts for approximately 0.5% of total cases of ID [442, 443]. ID is a prime example for genetically heterogeneous disorders with the number of genes found to be associated with ID reaching four digits [52, 435, 444]. The functions of the genes associated with ID are very diverse (e.g. chromatin remodelling, protein modification, differentiation of neural cells of the nervous system, centrosome function) with synapse formation and transmission related pathways being very common [202, 445]. See section 9.6.4 for a table of all identified ARID causal genes (Table 10.12), the phenotypes caused and references (also Table 10.11 in section 9.6.3 contains a list of all ARID causal/associated genes derived from GeneCards [411]).

5.2. Hypothesis

The hypothesis of this analysis was that the ID which segregates in the family analysed in this chapter was due to the presence of an autosomal recessive mutation in an autozygous region in the affected offspring (therefore the mutation will be in a homozygous state in all affected and not in the unaffected).

5.3. Aims and Objectives

The aim of this chapter was to identify the autosomal recessive intellectual disability causal region (and subsequently gene and variant) in a large consanguineous family with 6 affected members. The causal region was identified using autozygosity mapping.

5.4. Methods

5.4.1. Constructing Family Pedigree

From information gathered in meetings with individuals 21-24, a family pedigree was drawn (Figure 5.1). Multiple consanguineous marriages have occurred (between 15 and 16 who are second cousins once removed; 17 and 18 who are first cousins; 21

and 22 who are first cousins; and 23 and 24 who are first cousins) within the pedigree as a whole. This occurrence of multiple loops of consanguinity will increase the probability of an allele being IBD in the offspring, which is denoted as the F value here.

The expected and observed F values will be calculated using methods stated in section 2.4.6.

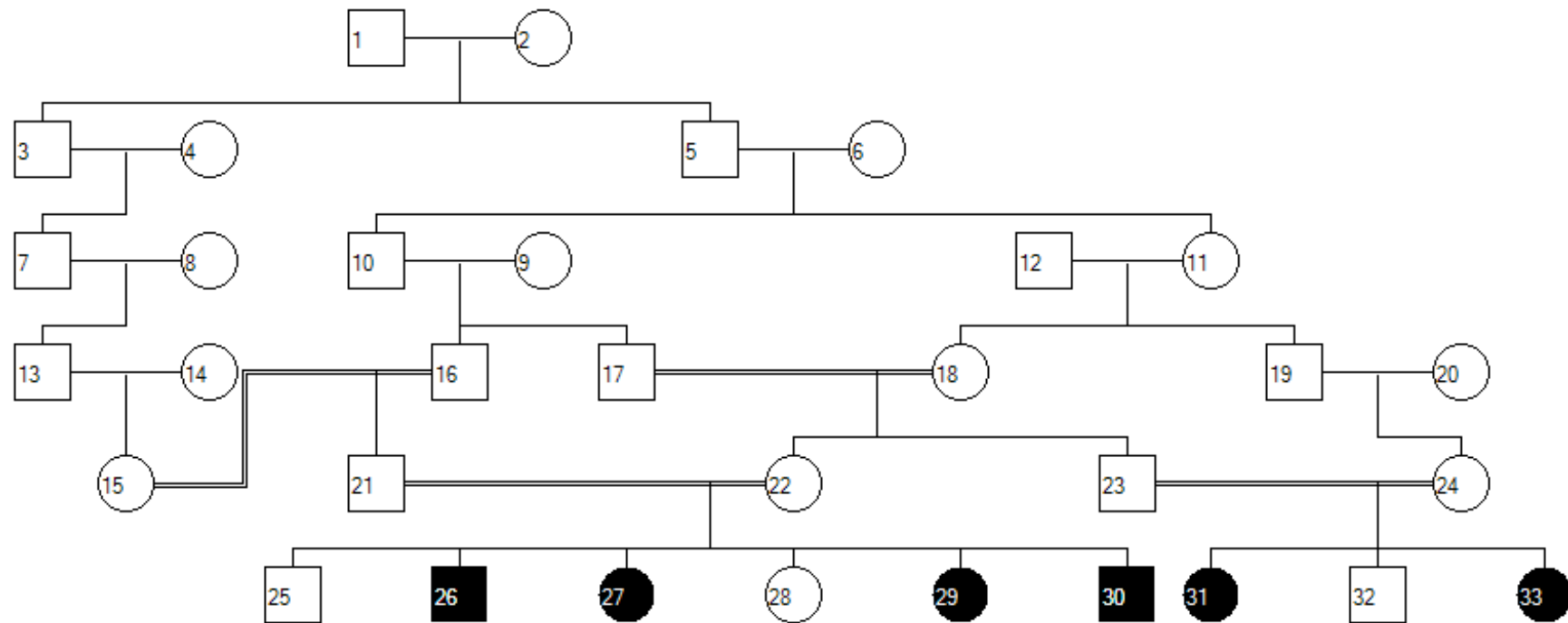


Figure 5.1 Whole family pedigree of participating family. Square: Male, Circle: Female. Individuals filled with black are affected. AutoSNPa was used to create the pedigree and export the image [297].

5.4.2. Determining LRoH from SNP array data

Genotyping data for 13 (6 affected and 5 unaffected) members of the family (depicted as 21-33 in Figure 5.1) was obtained from the direct-to-consumer company 23andme (reference human assembly build 36). The same 960566 SNPs (930342 autosomal SNPs, 26000 SNPs in chromosome X, 1765 SNPs in chromosome Y and 2459 SNPs in mtDNA) were genotyped throughout the genome of all participants. The text file was then anonymised (even though data processing is carried out in local server and not transferred across servers) and the LRoH of each individual was determined using David Pike's method (designed especially for 23andme SNP chip array data, available online at: <http://www.math.mun.ca/~dapike/FF23utils/roh.php>) setting required parameters as:

- (i) Report LRoHs of length of at least 150 consecutive SNPs
- (ii) Treat as homozygous any heterozygous SNP that is at least 150 SNPs away from its nearest heterozygous SNP

The same procedure was also carried out using Plink (23andme data can be converted to Plink format using Convert_23andme_to_plink.pl Perl script available in section 9.8) and AutoZplotter (Plink format can be converted to VCF format using Convert_plink_to_vcf.py available in section 9.8). To view the outputs in IGVtools suite (as a BED file) [284], the outputs were then copied to MS Excel and text was divided into columns, setting space as a delimiter. All columns except 'chr no' (e.g. 1,2), the 'start position' and the 'end position' (e.g. 7076127 and 71680128) were removed. The term 'chr' (in column A) was inserted before every 'chr no' (in column B) column then the two were merged in a different column using the formula:

$$=(A1\&B1)$$

Then the formula was altered accordingly as the row numbers increased from 1 all the way to the end (e.g. "=(A10\&B10)"), using the pull down option (provided by

MS Excel). The columns were then rearranged to fit the format of a BED file (detailed information can be found at: <http://genome.ucsc.edu/FAQ/FAQformat.html>):

Example:

```
Chr1 6555538      7253934
Chr1 70761270     71680128
```

The resulting file was then exported as a tab delimited file. The file was converted to '.bed' extension (by changing .txt to .bed). IGV was executed using the command:

```
java -Xmx1500m -jar path/to/igvtools.jar
```

and the bed files of each individual were uploaded to IGV for viewing.

5.4.3. Brute-force pinpointing of causal region(s)

Besides manually reviewing the identified LRoHs in section 5.4.2 then manually ruling out regions due to their presence in the unaffected members, a brute-force SNP by SNP approach was also carried out to determine whether there were any SNPs which fit all the criteria to be 'causal' (i.e. homozygous in all affected and not in unaffected). This would enable identifying very short regions (< 1 Mbp) which may evade available autozygosity mapping tools through the analysis of LRoHs. Every SNP (and the region it was tagging) was treated as a potential candidate and not just the ones which were in autozygous regions. For this the following commands were used (see section 9.8 for the files/scripts in **red** below):

- (i) To output all homozygous mutations in unaffected family members:

```
grep -f Homozygote_alleles.txt genome_patient1.txt >
patient1_homozygote.txt
```

- (ii) To remove non-causal homozygous mutations from proband*:

```
python Allele_remover.py
```

*change unaffected family members' names in python script accordingly until all non-causal SNPs deleted

- (iii) Finding alleles homozygous in all affected family members:

```
grep -x -f proband.txt patient2_homozygote.txt >
candidate_mutations_patient2.txt
```

then

```
grep -x -f proband-patient2.txt patient3_homozygote.txt >
candidate_mutations_patient3.txt
```

until the last affected member in the family...

5.4.4. Haplotype Phasing

As there is a complex consanguineous union loop within the family, it can be interesting to find out from which branches the mutation was passed down from and/or how long the inherited ancestral chromosomal segments were. Although we only had the SNP array data of two generations, haplotype phasing may give us clues about the answer.

Haplotype phasing was carried out using the BEAGLE software [291] as it provided a user-friendly interface for conversion of different data formats and could handle the massive amounts SNP data which is produced by 23andme. For details on the commands used, see section 9.8.

5.4.5. Protein structure modelling

Once the p.V128M variant in *ADAT3* was identified as ARID causal by Alazami *et al* (harboured within the region identified here via autozygosity mapping), I used the Robetta software to predict the effect of the mutation (they reported) on the protein structure of ADAT3p - as described in section 4.4.9.

The amino acid sequences submitted for ADAT3p are available in section 9.6.1.

5.5. Results

5.5.1. *F* values

The calculated and observed *F* values for individuals 21-33 are as follows:

Expected F values:

Individual 21 = $0.0078125 (0.5^8 \times 2)$

Individuals 22 and 23 = $0.0625 (0.5^5 \times 2)$

Individuals 25 to 30 $\approx 0.0670 (0.5^5 \times 2 \times 1.0078125 \times 1.0625)$

Individuals 31, 32 and 33 $\approx 0.0664 (0.5^5 \times 2 \times 1.0625)$

All other individuals ≈ 0 (due to outbreeding of parents)

Observed F values

The F values are rounded to nearest thousandth where necessary:

Individual 21 = $0.067 (206602514/3.1 \times 10^9)$

Individual 22 = $0.08 (255480323/3.1 \times 10^9)$

Individual 23 = $0.07 (217481347/3.1 \times 10^9)$

Individual 24 = $0.028 (85580182/3.1 \times 10^9)$

Individual 25 = $0.127 (394177688/3.1 \times 10^9)$

Individual 26 = $0.118 (365155010/3.1 \times 10^9)$

Individual 27 = $0.14 (433757719/3.1 \times 10^9)$

Individual 28 = $0.135 (419163341/3.1 \times 10^9)$

Individual 29 = $0.129 (400920554/3.1 \times 10^9)$

Individual 30 = $0.10 (314825731/3.1 \times 10^9)$

Individual 31 = $0.10 (308528560/3.1 \times 10^9)$

Individual 32 = $0.08 (261422210/3.1 \times 10^9)$

Individual 33 = $0.108 (334847478/3.1 \times 10^9)$

5.5.2. Autozygosity mapping

Using David Pike's algorithm as stated in section 5.4.2, LRoHs were determined for each affected individual and their unaffected siblings (See Table 10.10 for all LRoHs identified in the six affected offspring). All LRoHs were manually viewed in IGV [284] and a region 1.5 Mb long on chromosome 19 was the only one where all the affected individuals had overlapping LRoHs. See Figure 5.2 below for alignment of LRoHs of each affected subject.

5.5.3. Brute force mapping

To ensure a non-biased analysis (i.e. taking non-LRoHs into account also), all SNPs were analysed on their own. The method in section 5.4.3 was followed. 45 SNPs were found to be fitting autosomal recessive mode of inheritance in the six affected individuals (Table 5.2). All of the SNPs which fitted the autosomal recessive mode of inheritance were located on chromosome 19 between positions 1858479 and 2608397.

SNP ID	Chr	Position	Genotype
rs2028261	19	1858479	Homozygous
rs7260336	19	1861106	Homozygous
rs8109179	19	1867392	Homozygous
rs11084919	19	1873599	Homozygous
rs3810415	19	1875653	Homozygous
rs6510629	19	1876389	Homozygous
rs8730	19	1876942	Homozygous
rs11669357	19	1903586	Homozygous
rs2041120	19	1912474	Homozygous
rs7254820	19	1937455	Homozygous
rs11669527	19	1938279	Homozygous
rs12609309	19	1940480	Homozygous
rs10416239	19	2009331	Homozygous
rs10404242	19	2010585	Homozygous
rs1004320	19	2033678	Homozygous
rs12609225	19	2051346	Homozygous
rs3803915	19	2111529	Homozygous
rs6510657	19	2251886	Homozygous
rs8111911	19	2257575	Homozygous
rs16991095	19	2261756	Homozygous
rs7251424	19	2272430	Homozygous
rs7254110	19	2356559	Homozygous
rs3848632	19	2360865	Homozygous
rs3848635	19	2361332	Homozygous
rs12463182	19	2366245	Homozygous
rs11881243	19	2371674	Homozygous
rs7343178	19	2375938	Homozygous
rs743578	19	2381232	Homozygous
rs4807263	19	2382272	Homozygous
rs1865111	19	2389119	Homozygous
rs2288943	19	2389884	Homozygous
rs4806849	19	2398215	Homozygous
rs13345846	19	2434713	Homozygous
rs7254861	19	2468875	Homozygous
rs16989558	19	2489668	Homozygous
rs7260635	19	2502470	Homozygous
rs4806871	19	2586059	Homozygous
rs11084950	19	2608397	Homozygous

Table 5.1 SNPs tagging the ARID causal region (in accordance with autosomal recessive mode of inheritance).

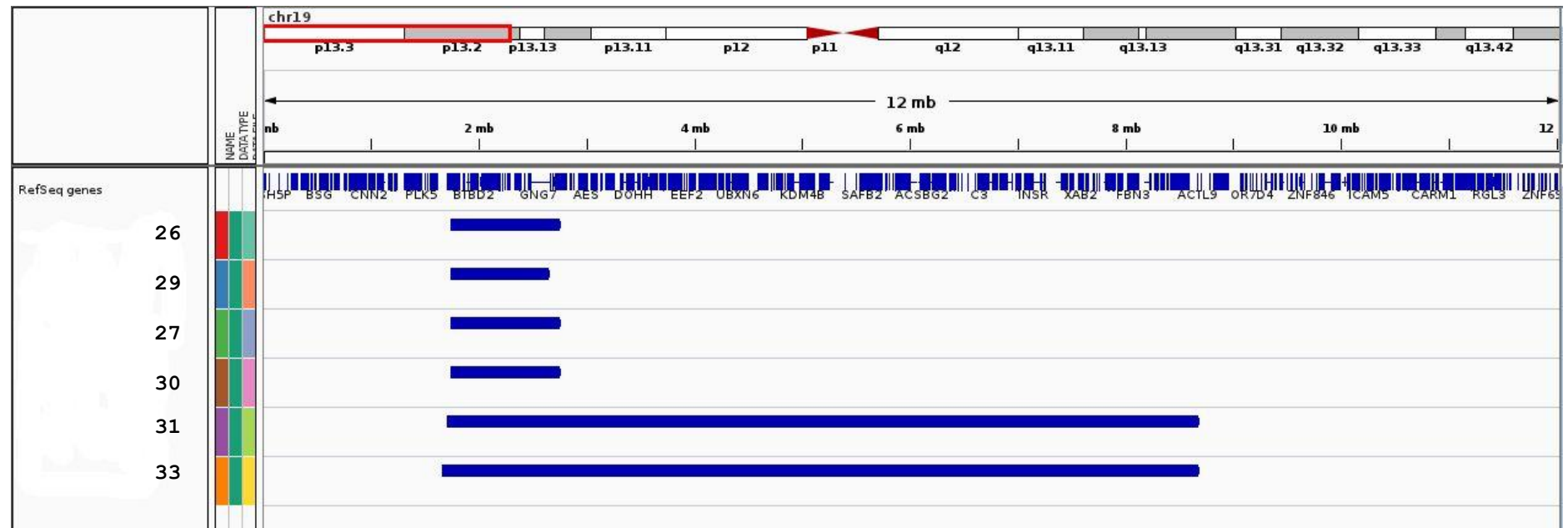


Figure 5.2 Using autozygosity mapping to pinpoint ARID causal loci. Individual IDs (in accordance with Figure 5.1) and their LROH at this chromosomal region are displayed on the left and right hand side of figure respectively. The depicted 1.5 Mb long region on chromosome 19 was the only one where all the affected individuals had overlapping LROHs. Image created using IGV [284].

5.5.4. Haplotype phasing

As a side analysis, individuals 21 and 22's haplotypes at the loci of the longest LRoH identified in the autozygosity mapping study within the whole family (i.e. between A₀ and B in Figure 5.3) were compared with individual 33's haplotype (Figure 5.1).

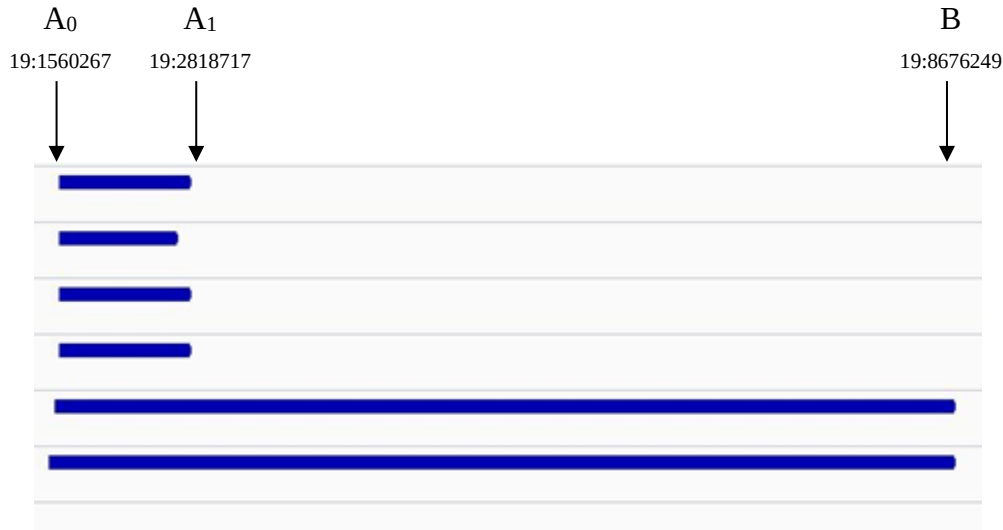


Figure 5.3 Setting LRoH boundaries between the affected individuals in the two related families. The genomic location of the SNPs can be found under each boundary label.

Before this analysis, a sensitivity analysis was carried out to check whether SNP array data at certain loci matched expected results. The results are as follows:

- (i) Haplotype between A₀ and A₁: Individual 21 = 22 = 26 = 27 = 29 = 30 (≥95% similarity)
- (ii) Haplotype between A₀ and B: Individual 23 = 24 = 31 = 33 (≥95% similarity)

After confirming SNP chip array reliability at the loci, the aforementioned analysis was carried out. Similarity between individual 21 and individual 33's haplotypes between A₀ and B was 10% more (95%) in relation to similarity between individual 22 and individual 33 (85%).

5.5.5. Modelling mutation effect to protein

Uploading the amino acid sequences stated in section 4.4.9 to the Robetta server yielded 5 different models for both the mutant (p.V128M) and the wild type ADAT3 protein. Figures 5.4 and 5.5 below depict the first model for both. The affected residue is labelled in both.

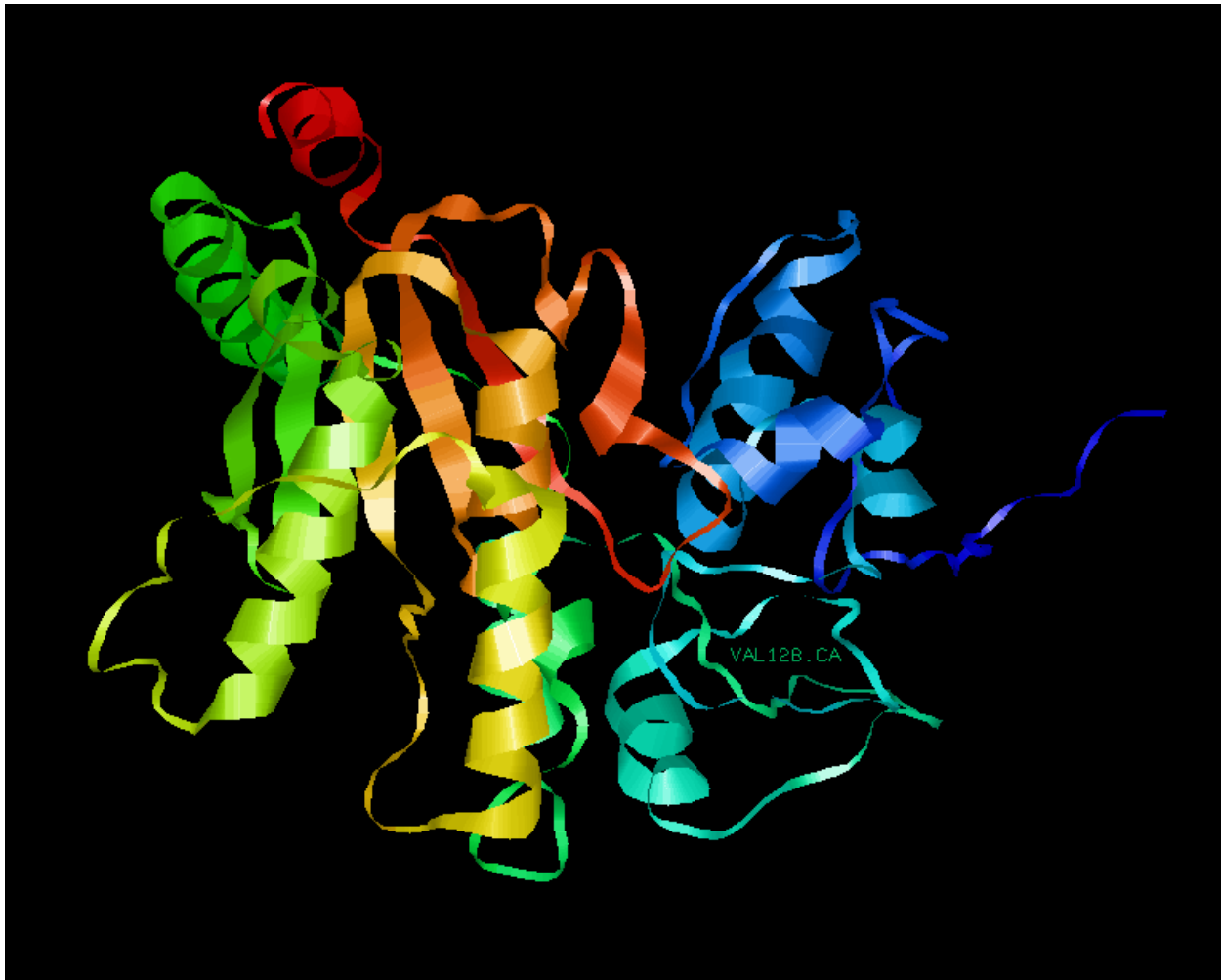


Figure 5.4 Protein structure of wild type ADAT3 protein. Only the first model is shown and the location of the mutation (p.V128M) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.

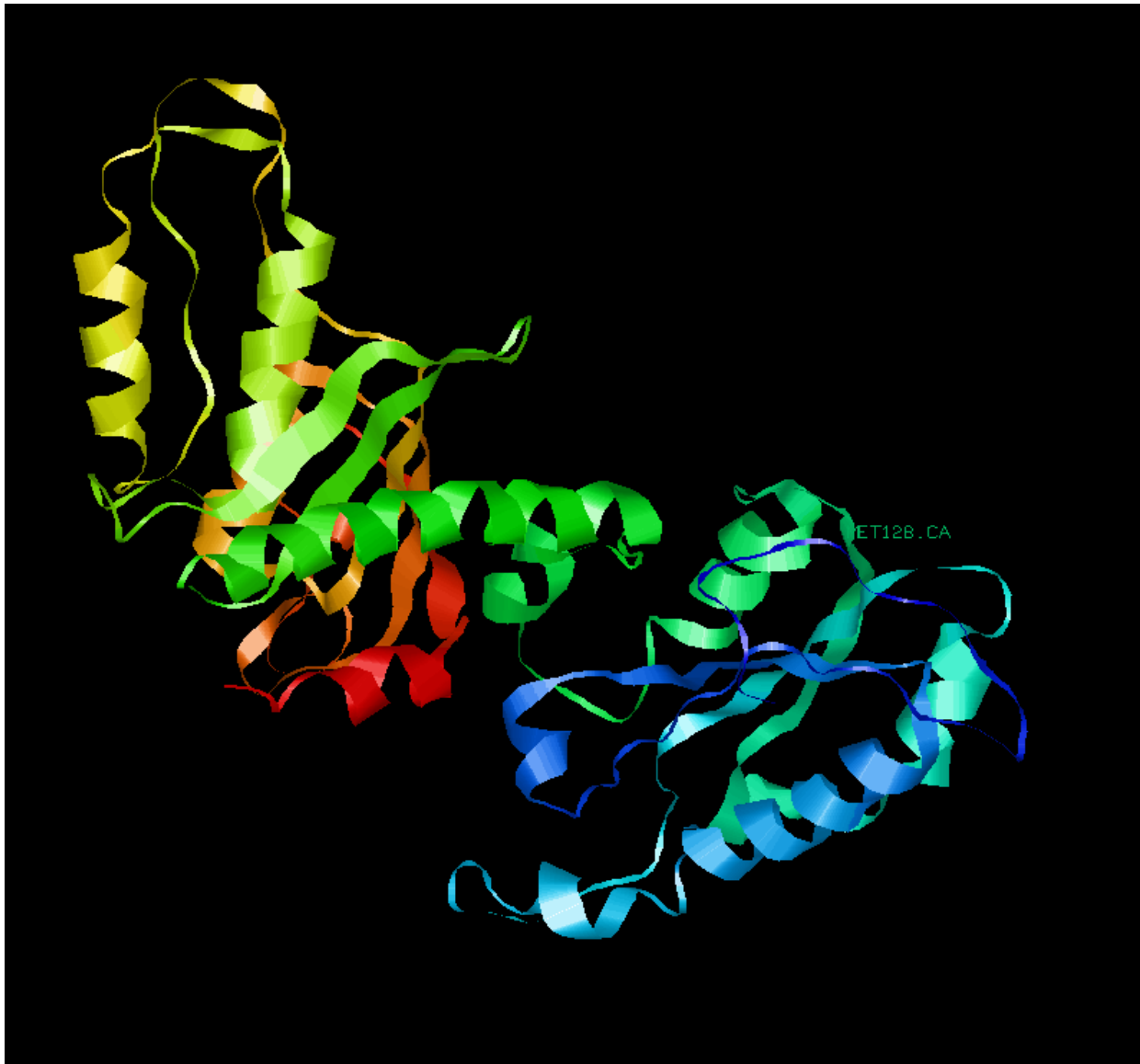


Figure 5.5 Protein structure of mutant ADAT3 protein. Only the first model is shown and the location of the mutation (p.V128M) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.

5.6. Discussion

Intellectual disability (ID) is a prime example for a genetically heterogeneous disorder with hundreds of genes already identified. Technological advances and decrease in costs in DNA sequencing and genomic microarray analyses have facilitated the rapid progress in gene identification that has been observed in the last few years. There is some difficulty in dissecting ID from other neurobehavioral and/or neuropsychiatric disorders (e.g. autism, schizophrenia) as they coexist in the same individual in a considerable proportion of cases. However despite all the

aforementioned advances in gene discovery related technologies and methods, a large proportion of cases of ID's aetiology still remains unknown [202].

Out of the genes that have been identified, many are *de novo* mutations and are very rare in the population thus the results still require replication in other families/individuals to be confirmed as causal. Validation of these novel genes/variants is vital and presumably for this to happen at an efficient scale, large-scale studies are required to discover new genes, replicate previously identified ones as well as estimate the penetrance of these mutations. Also through these analyses, the biological pathways related to ID will become clearer (e.g. major genes, modifiers), and monogenic and multifactorial aetiologies of ID will be identified. Animal models can also be useful where replication is not possible in humans.

As the main concern of this thesis was consanguinity and consanguineous individuals, autosomal recessive form of ID was analysed in a large consanguineous family with 6 affected children. The high number of affected offspring and no affected parents indicated a common autosomal recessive mutation segregating within the family. As is often the case, autozygous regions were analysed and a region on chromosome 19 was identified to be associated.

ADAT3 (and the p.V128M variant) was then identified as an ARID causal gene (and variant) by Alazami *et al*, which falls in the region we identified but is not a well studied gene. Gene ontology annotations for the gene's functions include hydrolase activity (derived from www.genecards.org), however it is not clear how this causes the phenotypes observed here.

5.6.1. Information from SNP array data

Before application of autozygosity mapping to the SNP array chip data, *F* values for each individual was calculated and measured separately to validate pedigree and consanguinity of parents. Expected *F* values were lower than the observed *F* values in most individuals. This is partly expected because of additional endogamy and/or consanguinity which may have been missed due to lack of data from previous

generations. But presumably the observed F values can also be over-estimated due to some monomorphic SNPs in the Saudi population being picked up as 'LROH's (used as proxy for autozygous regions in this analysis). This is likely to be the case as the SNP chip arrays used by 23andme are designed mainly as a result of studies carried out in individuals with European descent thus their LD patterns (and therefore tag SNPs) are likely to differ in comparison to Arabic individuals.

The haplotype phasing analysis confirmed the high reliability of the SNP chip array data obtained from 23andme as there was over 95% similarity* over a long chromosomal region where the compared individuals are expected to match 100%. This information was then used to compare the parents (i.e. individuals 21 and 22) of one family with the affected offspring of the other (i.e. individual 33). Individual 21 was found to be more similar to individual 33 in relation to individual 22 at the compared loci, which suggests that individual 21's family inherited and passed down a larger segment of the ancestral haplotype which carried the causal mutation.

5.6.2. Addition to literature

In this study, a 1.5 Mb long region located on the small arm of chromosome 19 was found to be associated with ARID. The results from the brute force study was complementary to the autozygosity mapping results as all the SNPs which were in accordance with the autosomal recessive mode of inheritance of intellectual disability were all located on the same region of chromosome 19.

However the main addition to the literature comes from the fact that we have replicated another research group's findings who also went onto publish before us. Alazami *et al* [446] had identified the same region using more families than we had (8 families, including the family we analysed) and sequenced the region to identify the causal gene and variant (i.e. p.V128M) itself.

Comparison analysis of the mutant and wild type forms of the ADAT3 protein shows that the former has a more open 3D structure. This change in structure can

* A few mismatches are expected mostly due to no-calls and genotyping errors

alter the function of the protein especially if an active site has been affected due to the mutation.

5.7. Conclusions

In this chapter, I have presented work from a large family study with six affected children diagnosed with ARID. The causal region was identified using autozygosity mapping and mapped to a 1.5Mb long region in chromosome 19. The next step would have been to amplify the region using PCR and then sequence the amplicons to identify the causal variant. Another group located in Saudi Arabia have studied the same family and found the causal variant to be p.V128M in the *ADAT3* gene which is located within the region we identified. Therefore the findings here would serve as a replication of their results. *ADAT3* is a useful addition to the tens of previously identified ARID causal genes (see 9.6.4 for details on these genes).

CHAPTER 6. PROXY MOLECULAR DIAGNOSIS FROM WHOLE- EXOME DATA REVEALS PAPILLON-LÈFEVRE SYNDROME CAUSAL MUTATION

In this chapter I have introduced Papillon-Lèfevre syndrome (PLS) as a disease, focusing on what is known about PLS at current, and also presented the findings of this thesis. This study was initiated after information about a second disorder within one of the participating families (i.e. Family 1 in Chapter 4) was provided to us, and a proxy molecular diagnosis was carried out using previously available WES data from other siblings. The ethical, medical and research implications of this study was also discussed in this chapter.

6.1. Introduction

Papillon-Lèfevre syndrome (PLS, MIM# 245000) is an autosomal recessive disorder characterised by severe early onset periodontitis and palmoplantar hyperkeratosis, which consequently results in the premature loss of the primary and secondary dentitions [447]. As the name suggests, the disease was first described by Papillon and Lefevre in 1924 [448].

PLS is caused by mutations in *CTSC* which displays remarkably high allelic heterogeneity with over 70 mutations of all types (e.g. nonsense, missense, frameshifting, UTR variants) reported hitherto [448]. *CTSC* encodes the cathepsin C protein, a lysosomal exo-cysteine proteinase belonging to the peptidase C1 family

[448]. The protein was first characterized by Turk *et al* in immune and inflammatory cells as an activator of serine proteases [449].

6.2. Hypothesis

The hypothesis of this analysis is that PLS which is diagnosed in the two (PLS) affected individuals is due to the rare (heterozygous) variant in the *CTSC* gene of their PCD affected (PLS unaffected) sibling who was previously studied (see section 4.5.2 for details on family 1). This variant will be studied further in the PLS affected siblings to test this hypothesis.

6.3. Aims and Objectives

The aim of this chapter was to utilise the already available WES data to carry out a 'proxy molecular diagnosis' on siblings of a previously studied PCD patient who had later been diagnosed with PLS.

6.4. Methods

6.4.1. Ethics

Ethical approval was obtained as stated in section 2.2. The participating family was previously analysed for a PCD study [80, 217]; and this study was carried out after the family re-attended the clinic enquiring about the cause of PLS present in other siblings within the family.

For this study, parental consent was from the father, with both parents and their children attending clinic together.

6.4.2. Probability of proxy diagnosis

We have previously studied the family here for a Primary ciliary dyskinesia (PCD) study where we whole-exome sequenced the two PCD affected children (the causal variant remains unknown, see family 1 in section 4.5.2). However, family history

indicated that there were two other siblings diagnosed with PLS and the family requested a molecular diagnosis for this disease also. We reasoned that with a one-third chance of each PCD (PLS unaffected) sibling with whole-exome sequencing (WES) data being a non-carrier for a PLS causal mutation, we would have an 8/9 chance ($1/3 \times 1/3 = 1/9$ = chance of both **not** being a carrier) of identifying the PLS causal mutation (likely in *CTSC*) in a heterozygous state in at least one of the two available WES data.

6.4.3. Participants and Genetic Data Analysis

A male proband from a consanguineous family of Arabic descent with clinical features consistent with PLS including loss of primary teeth and nail dystrophy was analysed. Additionally, three siblings' (one affected and two unaffected, including the PCD affected) and the parents' blood samples were also collected for further analysis. DNA was extracted from peripheral blood samples using the QIAamp DNA Mini kit provided by QIAGEN (Catalogue No: 51304) using their protocol for "DNA Purification from Blood or Body Fluids". The exome of the PCD affected sibling was captured using the Agilent SureSelect Human All Exon 50M exon capture kit (Agilent Technologies Inc., Santa Clara, CA, 95051, USA) and WES data was obtained by subsequent sequencing using the Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA, 92122, USA). The Burrows-Wheeler Aligner (BWA) [247] software was used to align the reads to the latest human genome reference sequence (hg19), filtering out reads which have extensive low base quality (more than half of the bases which have a base quality of ≤ 5 , including no calls) and/or with a mapping score of zero. Picard (<http://picard.sourceforge.net>) was used to mark duplicated reads and the alignment results were generated in BAM format. Single nucleotide polymorphisms (SNPs) were called using SOAPsnp [248] and small insertion/deletion events (indel) were detected by SAMtools and GATK, and exported in VCF format [249-251]. VCF annotations were obtained from the Ensembl Variant Effect Predictor (VEP) [252] and ANNOVAR [253]. Predictions for missense mutations were obtained from FATHMM [242], SIFT (via VEP) [243], Polyphen-2 (via VEP) [244] and Condel (via VEP plugin) [257]. The *CTSC* gene was screened for

variants which are either rare (<0.1%) or absent in the 1000 Genomes project [208] and Exome Variant Server (EVS) [450].

6.4.4. PCR amplification and Sanger sequencing

Region specific primers (i.e. where the NM_001814.4:c.899G>A;p.(G300D) variant is located) were designed and PCR (annealing temperature: 49C°) was used to amplify a 220bp long region containing the c.899G>A;p.(G300D) variant in the parents and the siblings (Forward primer: 5'- AAGCTAAGAACAACCTTTCAGGG-3' and Reverse primer: 5'- TGGAGAATCAGTGCCTGTGTAG-3'). These amplicons were purified and subsequently sequenced using Sanger sequencing.

6.4.5. Mutation screening using ARMS-PCR

Local Population

DNA was extracted from 256 unrelated (and healthy) individuals living in Riyadh using methods described above. Four primers (Control forward primer: 5'- AACATGCAAAGAATAATGGAG-3', Common reverse primer: 5'- AGCTTCATCAGGGCTTCATTG-3', Mutant allele-specific primer: 5'- TTCATCTTCAGGCTGTGAACG-3' and Wild-type allele-specific primer: 5'- TTCATCTTCAGGCTGTGAACA-3') were designed (see Table 6.2 for details) and ARMS-PCR (annealing temperature: 47C°, see Gaunt *et al.*, 2001 for description of method [451]) was used to detect the presence of the c.899G>A;p.(G300D) variant in 256 unrelated individuals selected from the local population in Riyadh. Resulting PCR amplicons were then viewed using 96-well microplate array diagonal gel electrophoresis (MADGE) [246]. Nucleotide numbering system uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon (Met) as codon 1.

Family members

The above procedure was also repeated on the family members to ensure validity of the method and confirm variant status using a different method.

6.4.6. Protein structure modelling

The structure of the CTSC protein and the effect of the mutation (p.G300D) on the wild-type form were modelled as outlined in section 4.4.9. In summary, the amino acid sequence of the gene under analysis is retrieved from the UniProt database and piped to the Robetta server for Ginzu prediction (a protocol that attempts to determine the regions of a protein chain that are aligned to PDB templates with reasonable confidence), and then for complete structure prediction. The same procedure was repeated for the mutant version of the protein. The most likely model (as determined by Robetta) was presented here (the next 4 most likely models were presented in the appendices, section 9.5.6).

Amino acid sequences uploaded to the Robetta server are available in section 9.5.6.

6.5. Results

6.5.1. Whole-exome sequencing of PCD affected sibling

The results are the same as in section 4.5.2 as the same WES data was used (see individual 2).

6.5.2. Screening CTSC gene

Whole-exome sequencing of the PCD affected sibling had previously been carried out (although no mutation causal of PCD has yet been identified) and her CTSC gene was analysed in follow up to the PLS presentation of two of her siblings [447]. 8 single nucleotide variations (intronic variants: rs217116, rs217060, rs580743, rs217075, rs217076, rs217077; missense mutations: rs217086 and c.899G>A:p.(G300D)) and a single nucleotide insertion (rs11426721) were identified in the CTSC gene. All except c.899G>A:p.(G300D) had an MAF of over 7% in the 1000 Genomes Project and EVS (see Figure 6.1 for alignment of reads) which are too common to be causal of a rare Mendelian disease such as PLS. FATHMM (damaging, -3.06), SIFT (deleterious, 0.01), Polyphen (probably damaging, 0.998) and ConDel-2 (deleterious, 0.880) all

predicted the c.899G>A:p.(G300D) variant to be functionally disruptive. The mutation also resides in a highly conserved region represented by a (36-way eutherian mammals) high GERP score of 1285.8 (also see Table 6.1 for local alignment with other species) [310]. Searching the public mutation databases and literature on the variant showed that it was previously identified in a homozygous state by Zhang *et al* in a single Saudi Arabian proband [452] and the variant was present in HGMD (Public version, ID: CM002939) and PhenCode (ID: CTSCbase_D0022:g.44271G>A) [298]. This provided strong evidence that this was the likely causal variant in the two PLS siblings. Thus the region containing the variant was amplified and sequenced using Sanger sequencing in both PLS affected siblings and the parents to confirm their status. In accordance with autosomal recessive mode of inheritance of PLS, the parents were heterozygous and the affected subjects were homozygous (Figure 6.2). The other PLS unaffected sibling was homozygous for the wild type allele. ARMS-PCR was also used in all family members to establish mutation status (Figure 6.3).

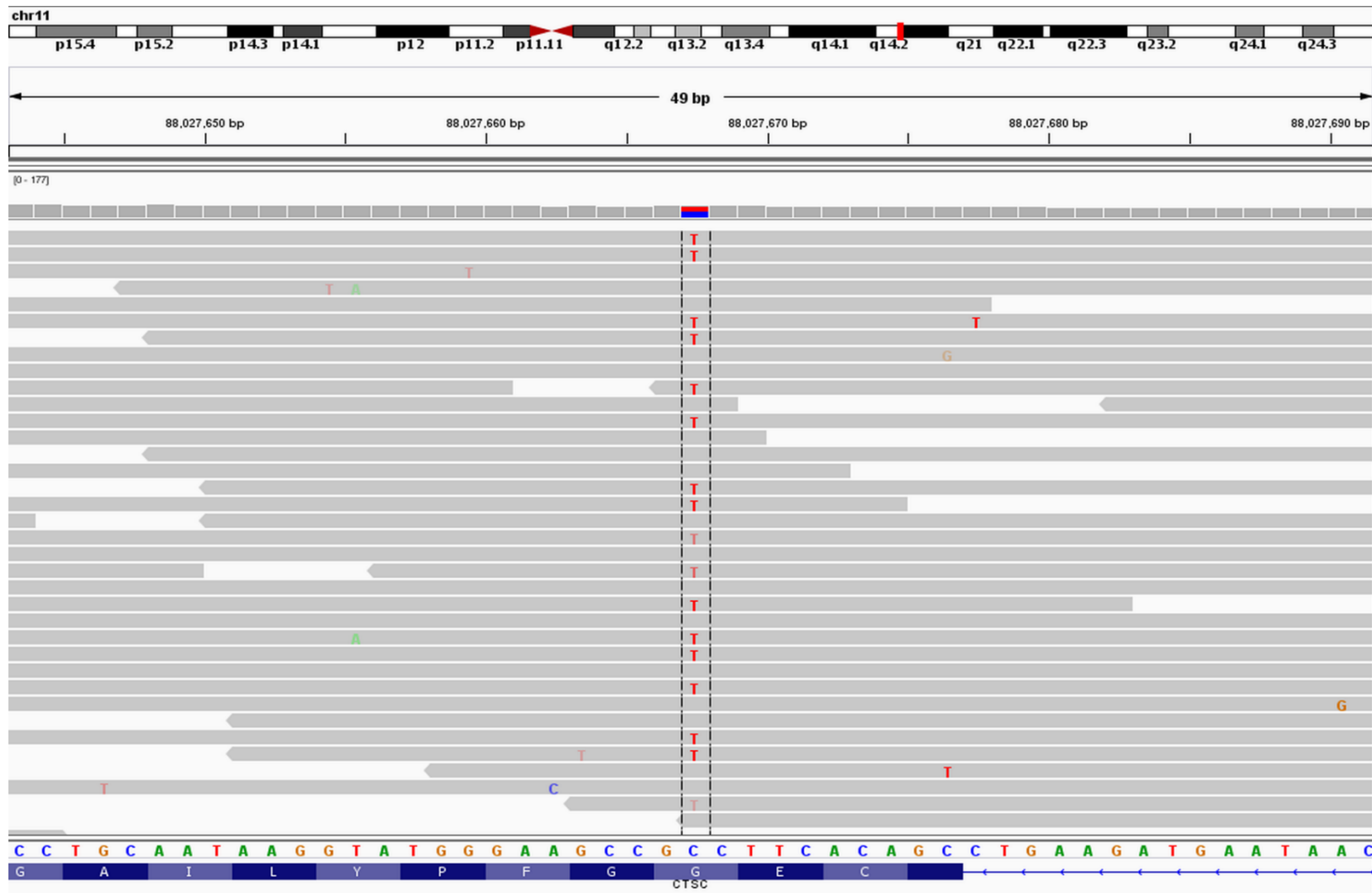


Figure 6.1 Reads (from PCD affected sibling) mapped to the human reference genome hg19 at the p.(G300D) mutation loci. The read depth is 46 (not all shown due to space restrictions) with twenty six of the reads having a G at the loci and twenty having an A. The image was created using IGV [24]. NB: The CTSC gene is oriented the reverse strand, therefore the codon change p.G300D (GGC>GAC) is exhibiting as C>T.

Start	Amino acid sequence	End	Entry	Entry Name	Organism
296	QGCE G GFPYLIAGKYAQDFGLVEE	319	<u>P53634</u>	CATC_HUMAN	<i>Homo sapiens</i> (Human)
295	QGCD G GFPYLIAGKYAQDFGVVEE	318	<u>P80067</u>	CATC_RAT	<i>Rattus norvegicus</i> (Rat)
269	QGCE G GFPYLIAGKYAQDFGLVEE	292	<u>O97578</u>	CATC_CANFA	<i>Canis familiaris</i> (Dog)
296	QGCE G GFPYLIAGKYAQDFGLVEE	319	<u>F1N455</u>	F1N455_BOVIN	<i>Bos taurus</i> (Bovine)
296	QGCD G GFPYLIAGKYAQDFGLVEE	319	<u>M3W9M0</u>	M3W9M0_FELCA	<i>Felis catus</i> (Cat)
295	QGCD G GFPYLIAGKYAQDFGVVEE	318	<u>P97821</u>	CATC_MOUSE	<i>Mus musculus</i> (Mouse)
289	QGCD G GFPYLI-GKYIQDFGIVEE	312	<u>Q6P2V1</u>	Q6P2V1_DANRE	<i>Danio rerio</i> (Zebrafish)
291	QGCE G GFPYLIAGKYVSDYGIVEE	314	<u>F7E2G8</u>	F7E2G8_XENTR	<i>Xenopus tropicalis</i> (Western clawed frog)
296	QGCE G GFPYLIAGKYAQDFGLVEE	319	<u>H2Q4I9</u>	H2Q4I9_PANTR	<i>Pan troglodytes</i> (Chimpanzee)
296	QGCE G GFPYLIGGKYAQDFGLVEE	319	<u>G3VM46</u>	G3VM46_SARHA	<i>Sarcophilus harrisii</i> (Tasmanian devil)
242	QGCD G GFPYLIAGKYTQDFGVVEE	265	<u>F7F2M7</u>	F7F2M7_ORNAN	<i>Ornithorhynchus anatinus</i> (Duckbill platypus)
296	QGCN G GFPYLIAGKYAQDFGLVEE	319	<u>G1T7L0</u>	G1T7L0_RABIT	<i>Oryctolagus cuniculus</i> (Rabbit)
296	QGCE G GFPYLIAGKYAQDFGLVEE	319	<u>G3SMJ4</u>	G3SMJ4_LOXAF	<i>Loxodonta africana</i> (African elephant)
296	QGC A GFPYLIAGKYAQDFGLVEE	319	<u>F1STR1</u>	F1STR1_PIG	<i>Sus scrofa</i> (Pig)
300	QGCE G GFPYLVAGKYAQDFGVIEE	323	<u>T2MEP6</u>	T2MEP6_HYDVU	<i>Hydra vulgaris</i> (Hydra attenuata)
---	QGC* G GFPYL**GKY**D*G**EE				

Table 6.1 Local sequence alignment containing the mutated residue from multiple alignment of the *CTSC* gene in different species. The highlighted Glycine (G) residue is found to be highly conserved across many species which have a homologue of the human *CTSC* gene. The alignment was carried out using the Uniprot website's Blast and Align functions (<http://www.uniprot.org>). The final row shows there is diversity in adjacent and nearby residues (represented by *), however the G residue is highly conserved.

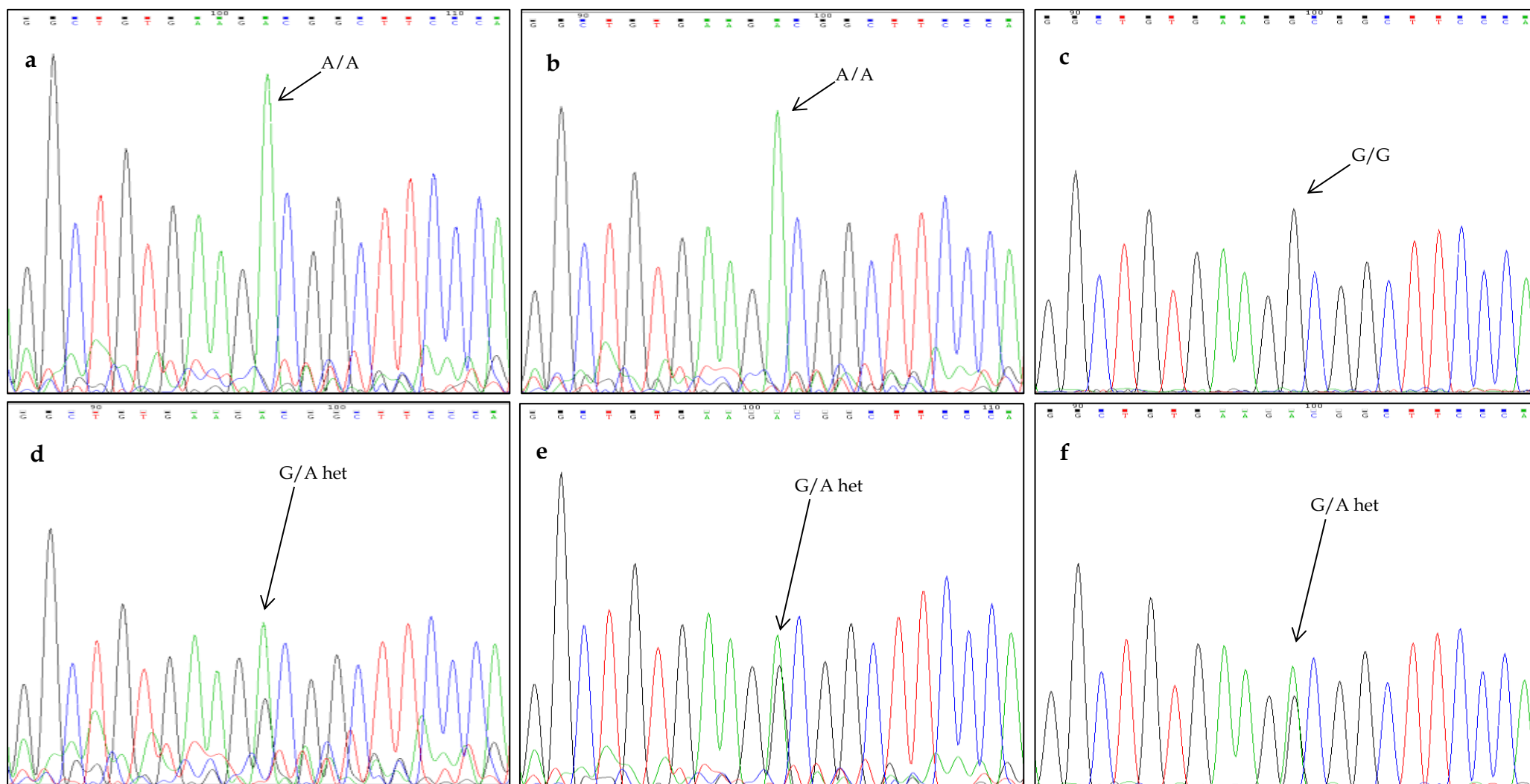


Figure 6.2 Confirmation of variant status in other family members using Sanger sequencing: (a) Male proband (b) Affected brother (c) Unaffected sister (d) Father (e) Mother (f) PCD affected (and PLS unaffected) sibling whose WES data was available.

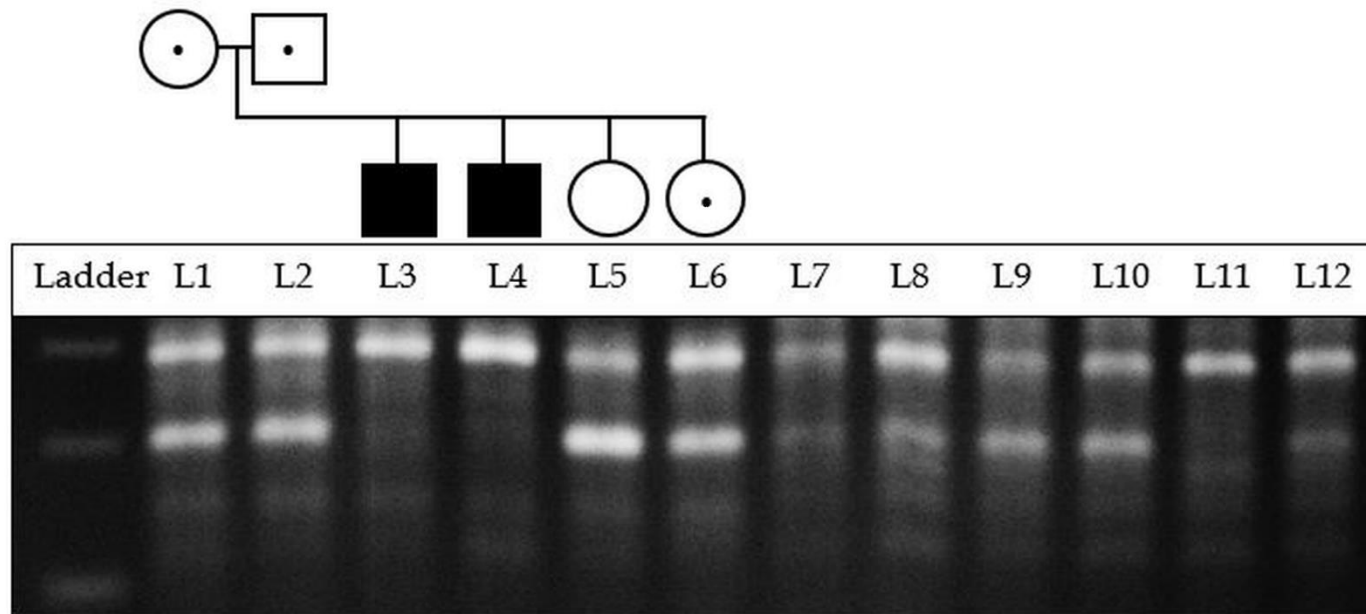


Figure 6.3 Validation of NM_001814.4:c.899G>A:p.(G300D) in all family members using ARMS-PCR. For primers, see Table 6.2. L1-L6: using AS primer for wild type. L7-L12: using AS primer for mutant. L1/7: Mother. L2/8: Father. L3/9: PLS Proband. L4/10: PLS Affected brother. L5/11: Unaffected sibling – homozygous for *CTSC* wild type allele. L6/12: PCD affected sibling who is a carrier for PLS. Ladder's three bands are 100bp (bottom), 200bp and 300bp (top).

6.5.3. Cathepsin C protein structure prediction

Uploading the amino acid sequences stated in section 6.4.6 to the Robetta server yielded 5 different models for both the mutant and the wild type CTSC protein. Figures 6.4 and 6.5 below depict the first model for both. The affected residue is labelled in both.



Figure 6.4 Protein structure of wild type CTSC protein. Only the first model is shown and the location of the mutation (p.G300D) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.



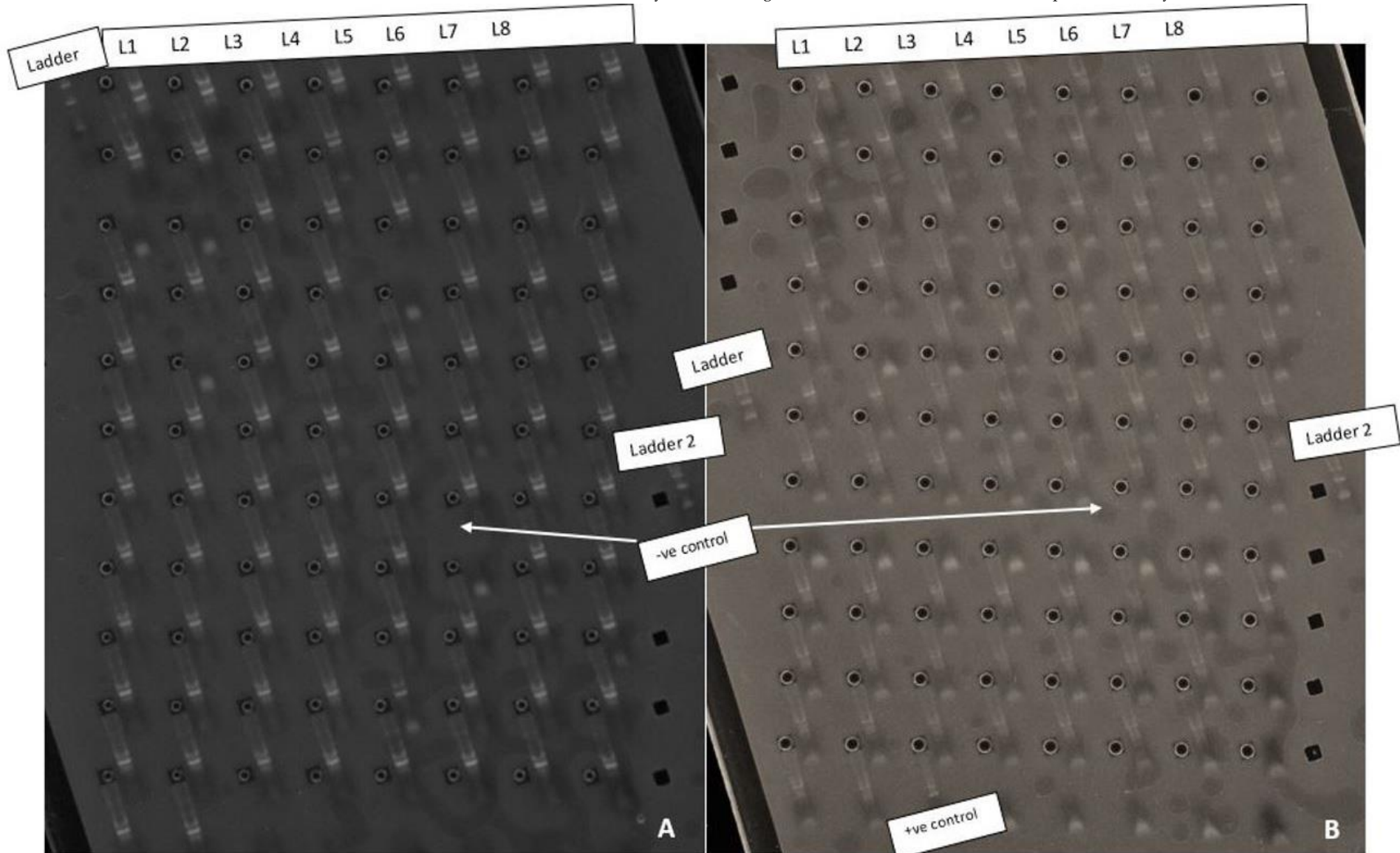
Figure 6.5 Protein structure of mutant CTSC protein (p.G300D). Only the first model is shown and the location of the mutation (p.G300D) is presented. For the other 4 models predicted, see section 9.5.6. The image was created using RasMol.

6.5.4. Screening for p.G300D variant in Riyadh, KSA

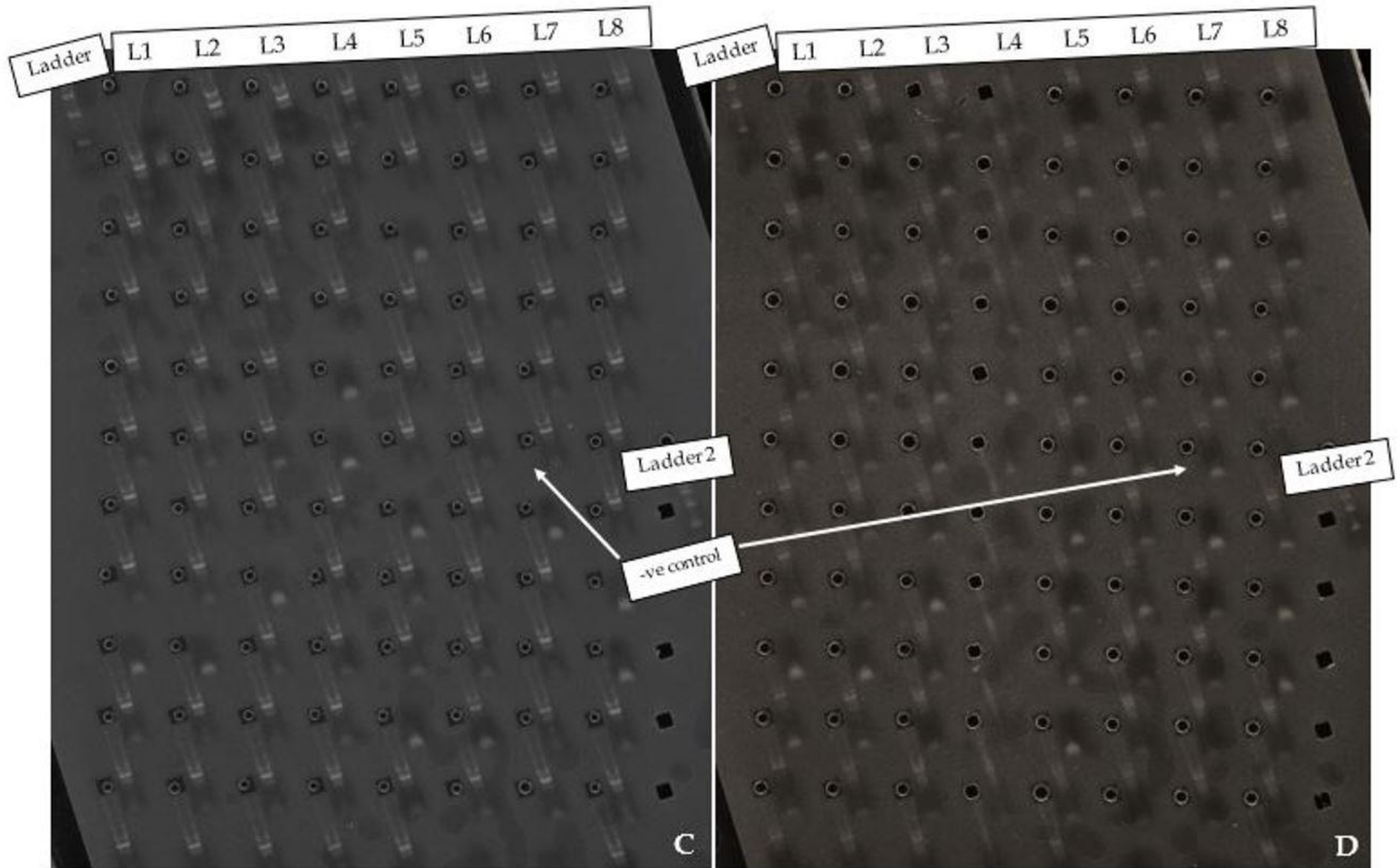
Allele-specific PCR amplicons (using primers in Table 6.2) from DNA from the 256 participants were separated using 96-well MADGE (procedure was repeated three times). None showed the 207bp band characteristic of the mutant allele, whereas the wild-type allele-specific band was present in all participants when wild type AS primer was used (Figures 6.6A-F). The results for all 6 family members are shown in Figure 6.3.

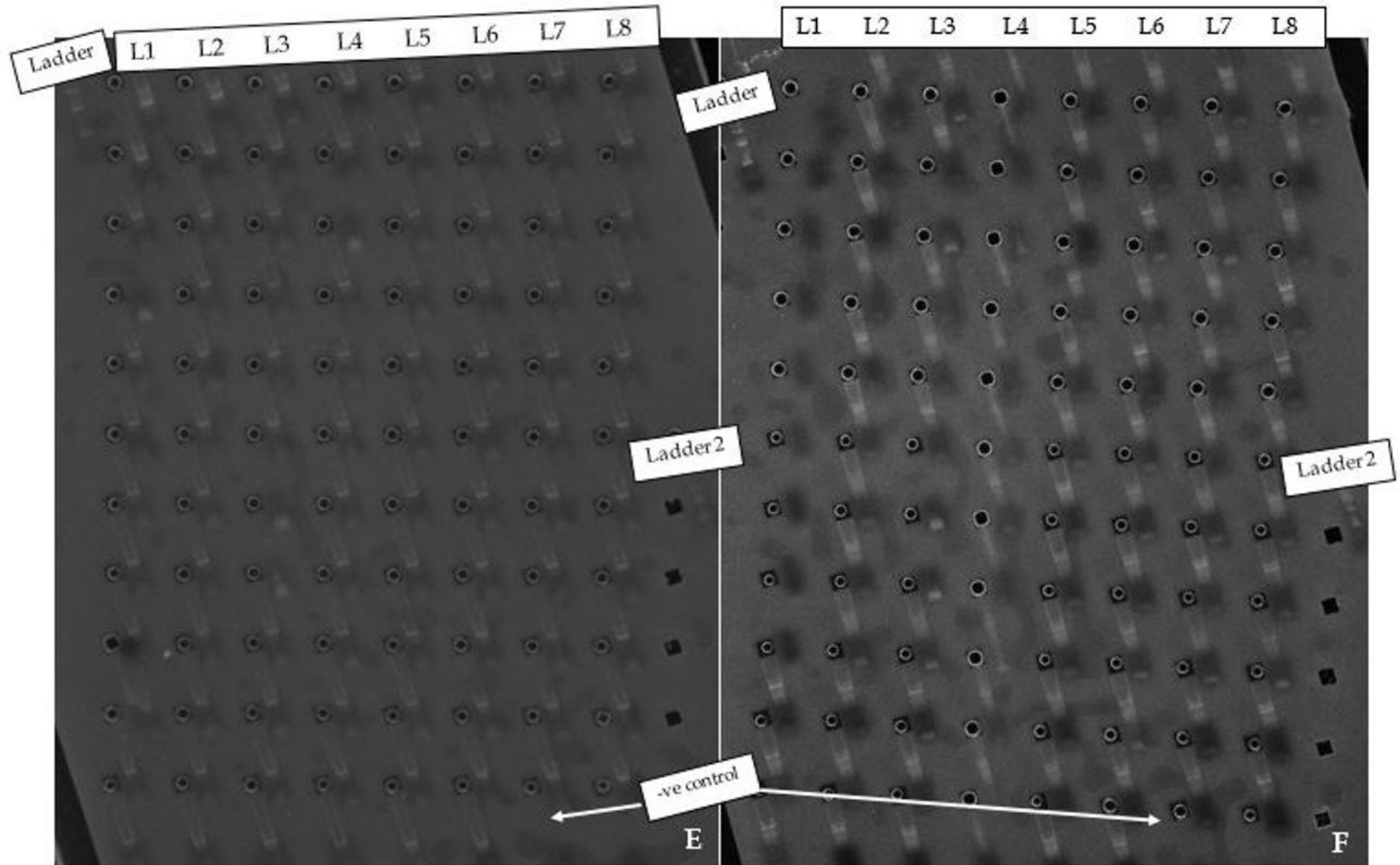
Genomic DNA (5' to 3')	Allele-specific primer – Mutant (5' to 3')	Allele-specific primer - Wild type (5' to 3')	Control primer (5' to 3')	Common primer (5' to 3')	Size of PCR amplicons
Wild type TTCATCTTCAGGCTGTGAAGG	TTCATCTTCAGGCTGTGAA A	TTCATCTTCAGGCTGTGAA G	AACATGCAAAGAATAATGGAG	AGCTTCATCAGGGCTTCATTG	Control-common = 291bp
Mutant TTCATCTTCAGGCTGTGAAGA					AS-common = 207bp

Table 6.2 Primers used in ARMS-PCR for genotyping the NM_001814.4:c.899G>A:p.(G300D) variant. The allele-specific (AS) primers have a mismatch at the -2 position (highlighted in green). The AS and control primers are used as forward primers in PCR; and common primer is a reverse primer.



Figures 6.6A-F (continues below) Screening the local population for the c.899G>A:p.(G300D) variant. 96-well MADGE images reveal that none of the 256 individuals have the causal allele. ARMS-PCR (A,C,E) using wild type primers (B,D,F) using AS primers. A, C and E are complementary to B, D, and F respectively. Ladder's three bands are 100bp (bottom), 200bp and 300bp (top).





6.6. Discussion

The *CTSC* gene displays high allelic heterogeneity and over 70 variants have been shown to cause Papillon-Lèfevre syndrome [448]. The c.899G>A;p.(G300D) variant is one of those, previously being reported in a single proband by Zhang *et al* [452]. Our findings follow up their paper as we have replicated their results, confirming the highly penetrant nature of the variant, and found that the prevalence of the variant in Riyadh, Saudi Arabia is rare (0/512 of control chromosomes analysed). We also present a straightforward and cost-effective assay to test for this mutation.

The variant was identified in a previously whole-exome sequenced and PLS unaffected sibling of the proband which shows how additional inferences can be made from WES (i.e. proxy molecular diagnoses). Thus, where WES (or whole genome sequencing) data is available and consent is given, it can be a pragmatic choice to screen for known mutations using databases such as HGMD (Public and Paid versions available), PhenCode (Public) and ClinVar (Public).

However, there are ethical issues surrounding incidental findings [453, 454]. WES data can be a source for these findings as it provides a pool of all detected variants in *all* genes. Therefore informed consent and abiding by the consent obtained is crucial (see reference [454] for a discussion on the matter). Our finding however, was not incidental and the study was carried out *only* after the family had attended the clinic with a second disorder (i.e. PLS) and gave consent for the subsequent analysis. We did not screen the family's previously available WES data other than for previously known/suspected PCD causal variants (in accordance with previous consent) before we were given further consent to search for the PLS causal variant. The *CTSC* gene was then screened using the available WES data and a missense variant which was previously reported as PLS causal was identified in a heterozygous state in one of the PCD affected siblings [452]. This then enabled us to make a proxy molecular diagnosis and confirm the variant's homozygosity status in the PLS affected siblings.

Our study highlights the wider and longer-term value of sequence data in the context of family history and additional clinical data. If it is stored and easy to query,

it provides considerable potential for future diagnostics within families at minimal additional cost. In this example, once the PLS was diagnosed in the proband it took only a few minutes before the causal variant was identified in the PCD affected sibling whose WES data was available - saving considerable time, effort and cost.

Comparison analysis of the mutant and wild type forms of the CTSC protein shows that the former has an altered 3D structure and has folded differently. This change in structure can alter the function of the protein.

6.6.1. On therapy and cures

PLS does not have a cure – like most inherited skin diseases, with little progress in developing effective and specific treatments [455]. However PLS is not a life threatening skin disease* where lost teeth can be replaced via dental implants (e.g. construct complete denture), and the visible dry scaly patches can be covered up with makeup, nail polish (where applicable) and/or regular skin care (e.g. topical keratolytics, oral retinoids). Most PLS patients live normal lives with typical life expectancies. Details on treatment information can be found at: <http://emedicine.medscape.com/>.

6.6.2. Addition to literature

Analysing whole-exome sequencing data from an Autosomal recessive congenital ichthyosis (ARCI) affected patient, Takeichi *et al* helped with the clinical diagnoses of ARCI in this patient while also identifying a Walker-Warburg syndrome causal variant in her in a heterozygous state which affected two of her siblings (one deceased) [456]. Their findings presented the additional information which could be gained from WES data. However their use of the WES data was in the direction of mutation-to-phenotype, not phenotype-to-mutation which was the case in ours.

The p.G300D variant identified in CTSC in this study was previously identified by Zhang *et al* [452]. However our study provides evidence for the highly penetrant

* Unlike xeroderma pigmentosum and epidermolysis bullosa for example

nature of the variant as both siblings who were affected by PLS were homozygous for the variant.

6.6.3. Considerations

The identification of the PLS causal variant is not the only significant contribution of this study. It becomes more noteworthy because the ethical, medical and research implications of this study will apply to all studies which deal with data obtained from consanguineous families.

6.7. Conclusions

In this chapter I presented work displaying the additional inferences which can be made with WES data. A missense mutation in *CTSC* was identified in a heterozygous state in a previously studied PCD patient. However, after two of her siblings were affected with PLS, her WES data was analysed again and a proxy diagnosis was made. The variant was then confirmed in the two affected siblings who fitted the autosomal recessive mode of inheritance that PLS follows.

CHAPTER 7. IMPORTANCE OF CONSANGUINEOUS POPULATIONS TO CHARACTERIZATION OF HUMAN GENE FUNCTION

In this chapter, I (with guidance from Prof. Ian N.M. Day at the University of Bristol) have put forward arguments about the importance of studying consanguineous populations as a whole (i.e. including unaffected individuals) rather than the traditional ‘cherry-picking’ of families with clinical disease phenotypes. I have used the analogy of reverse genetics (i.e. genotype to phenotype) studies carried out in model organisms throughout the chapter to better depict the additional inferences which can be gained by studying consanguineous populations as whole.

This chapter also contained a review (also see reference [457]) of the relevant literature with regards to the importance of consanguineous populations to human genetics (with the addition of my ideas on the respective subject matter).

7.1. Introduction

Consanguineous offspring have elevated levels of homozygosity. Autozygous stretches within their genome is likely to harbour loss of function (LoF) mutations which will lead to complete inactivation/dysfunction of genes. Studying consanguineous offspring with clinical phenotypes has been very useful for identifying disease causal mutations. However at current, most of the genes in the human genome have no disorder associated with them and/or have unknown function. This is presumably mostly due to not observing homozygous LoF variants

in outbred populations which are the main focus of large sequencing projects. However another reason may be that many genes in the genome – even when completely ‘knocked out’, do not cause a distinct and/or defined disease phenotype. Here I discuss the benefits and implications of studying consanguineous populations as a whole, as opposed to the traditional approach of analysing a subset of consanguineous families and/or individuals with disease. I suggest that studying consanguineous populations can speed up the characterisation of novel gene functions as well as indicating non-essential genes and/or regions in the human genome. I also suggest designing a SNV array to make the process more efficient.

In this chapter my aim was to present a theoretical cohort study with the primary aim of maximising the number of characterised ‘complete’ gene inactivations in the human genome. Many previously unknown genetic disorders (especially Mendelian disorders) have first been identified in consanguineous families [52, 133]; and many are still unique to consanguineous populations. This is due to the causal mutation(s) also being unique to regions where consanguinity levels are high; or the frequency of the causal allele can be extremely low in other outbreeding populations thus never getting the chance meet its counterpart and express its effect in a homozygous state.

Autozygosity mapping has proven to be a powerful technique for unearthing autosomal recessive disease causal mutations in consanguineous offspring [296]. However even after decades of studying consanguineous families with disorders, many genes in the genome still do not have a clinical phenotype associated with them. This could be due to most genes not causing a distinct/defined phenotype, which are the main focus of genetic association (and linkage) studies (especially early-onset disorders are studied). Therefore a paradigm shift is required in order to discover the function of the remaining genes. All genes need to be observed when completely ‘knocked out’ (i.e. become completely dysfunctional/inactivated by homozygous loss of function mutations, LoF) similar to reverse genetics* studies carried out in model organisms in order to understand better their functions.

* Where a gene is ‘knocked out’ and its consequent effects are observed

Consanguineous unions – albeit no way near – are the closest humans get to these types of studies. However it is clear that all the answers do not lie only in the consanguineous offspring with disorders. Consanguineous offspring without distinct clinical phenotypes should also be analysed to observe which genes have homozygous LoF mutations in them. This can then shed light into the function of these genes as these individuals can then be followed up through cohort studies (to see what the long term effects are, if any) and/or molecular studies (to observe any subtle differences e.g. change in other genes' expression). It may be that the gene is dispensable, thus is on its way to becoming a pseudogene. 'Knocking out' certain genes may even have protective effects against some disorders/diseases. In this chapter, my aims were to put forward a few arguments for undertaking a gene-centric approach rather than the traditional disorder-based approach when analysing consanguineous populations. I have also made suggestions about which consanguineous populations are most suitable for these types of analyses.

There are over 7500 disorders with known and/or suspected Mendelian basis, however 4473 have had their molecular basis determined (from OMIM; statistics true as of 23-06-15) [23, 52]. However, over a half of these are caused by autosomal dominant mutations (thus causing autosomal dominant disorders), and the rest are autosomal recessive or X-linked. Thus we have not observed the homozygous effects of mutations causing loss of function for over 10000 genes (up to 15000, excluding ~2000 genes located in sex chromosomes and the commonly 'knocked out' autosomal genes such as olfactory receptor genes). This is presumably mainly due to selecting on disease phenotypes in humans. My 'phenotypic ascertainment' claim is backed up by the high proportion of autosomal dominant disorders in comparison to autosomal recessive disorders. The Hardy-Weinberg (H-W) equation (see Equation 7.1) predicts that in an outbred population the proportion of heterozygotes (i.e. $2pq$) will be considerably higher than homozygotes of the causal variant ($2pq > q^2$, $2q$ for very low q), thus it is no surprise that when one ascertains for disease phenotypes, they will identify more autosomal dominant mutations [458]. I therefore suggest that sampling individuals randomly from a consanguineous population – as

opposed to only targeting families with disease, will evade this ascertainment bias and identify more homozygous LoF mutations in most, if not all genes – given ‘sufficient’ sample sizes. See Figure 7.1 for some examples of inferences which could be made from analysing autozygous regions in consanguineous offspring.

Also historically, discovering these gene inactivations in outbred populations (such as Western populations, which have been the focus of most genetic studies) would have required large enough families to determine the inheritance pattern making autosomal recessive mutations much harder to identify compared to autosomal dominant ones.

There are regions and sub-regions in the world where consanguineous unions are preferred for various socio-economic reasons with consanguinity levels reaching as high as 70% (as detailed in section 1.3.5) [185]. Despite the importance of consanguinity to genetic research, most of these populations are far from being thoroughly researched in genetic and sociological terms [86, 133]. Larger sequencing projects are required to make full use of these populations which is bound to serve human genetics immensely.

$$p^2 + 2pq + q^2 = 1$$

Equation 7.1 The Hardy-Weinberg equation (HWE). HWE assumes that allele frequencies in a population should not change under normal circumstances (e.g. random mating, no genetic drift) and provides an estimate for the frequency of heterozygotes and homozygotes. p : frequency of common (in this case wild-type) allele (allele: A), q : frequency of rare (i.e. disease causal) allele (allele: a) or $1-p$, p^2 : frequency of individuals with genotype AA, q^2 : frequency of individuals with genotype aa, $2pq$: frequency of individuals with allele Aa.

a) Homozygous deletion of Gene 1 in an individual with autosomal recessive intellectual disability (ARID)



b) Homozygous stop gain in Gene 2 which is a target for nonsense mediated decay (NMD) in a healthy individual



c) Homozygous stop gain towards end of Gene 3 which does not cause NMD and Gene 3 is known to cause an autosomal recessive disorder (e.g. Primary ciliary dyskinesia) from previous studies, however the individual is healthy with very mild symptoms



d) Gene 1 and Gene 21000 are both completely inactivated, however the individual is healthy although knocking out Gene 1 should have caused ARID (see Figure 1a)



Figure 7.1 Examples of inferences to be gained from autozygous regions in consanguineous offspring. (a) Homozygous LoF mutations in Gene 1 causes ARID (b) Gene 2 is likely to be a non-essential gene (i.e. dispensable). Subject should be followed up for late-onset effects (c) Although Gene 3 can cause PCD, coding region from stop gain to end of exon is not essential for correct functioning of gene, hence the unaffected subject (NB: mutation is not a target for NMD) (d) Although LoF mutations in Gene 1 causes ARID, concurrent inactivation of Gene 21000 due to NMD masks disease phenotypes indicating interaction between the two genes' products in the causal pathway (e.g. gain of function mutation at gene 1 could become dysfunctional by mutation at gene 21000). X: Stop gain. Ø: Deletion of whole gene. Position of stop gain within genes is for illustration purposes. This is not an exhaustive list of all the possible inferences which could be gained from studying consanguineous populations (e.g. identifying dispensable regions, proxy molecular diagnoses - see reference [459] for details on latter)

7.2. Hypothesis

The hypothesis of this study is that carrying out large-scale sequencing projects in consanguineous populations as a whole and establishing consanguineous collections (i.e. first cousin unions or closer) can speed up the characterisation of novel gene functions via observing their effects in a completely inactivated way as well as indicating non-essential genes and/or regions in the human genome.

7.3. Aims and Objectives

The aims of this chapter was to carry out a (non-systematic) review on the relevant literature with regards to the importance of consanguinity and consanguineous populations to human genetics, and to put forward arguments on the use of consanguineous populations 'as a whole' rather than only 'cherry-picking' certain families with disease. The chapter also aimed to demonstrate that carrying out large-scale sequencing projects in consanguineous populations as a whole is feasible and worth the costs in terms of characterising complete human knockouts. Suitable populations were also presented.

Not within the scope of this thesis, the ultimate aim will be to establish a resource which will consist of a dataset of whole exome of sequencing data obtained from a large number (>5000 participants) of offspring of consanguineous parents combined with basic anthropometric values (e.g. age, sex, weight, height) and health related phenotypes (e.g. medical records) – keeping the running costs at a minimum (in contrast, for an example of a 'phenotype rich' but also an 'expensive to run' cohort, see the Avon Longitudinal Studies of Parents and Children cohort, known as ALSPAC [460]). Once the database is established, the next aim would be to make the data available to research groups from around the world who will then initiate separately funded 'recall' projects to analyse the participants they are interested in (i.e. ones who possess the gene(s) of interest in completely inactivated form). The overall aim is to characterise as many novel gene functions as possible.

7.4. Natural human gene knock-outs in consanguineous populations

Empirical studies show that each individual possesses between ten to twenty (depending on ancestry*) rare mutations introducing premature nonsense codons [212], but in a heterozygous state. The number of mutations causing LoF will be increased with the addition of rare frameshifting indels (found to be between 8 and 17 indels, again depending on ancestry [212]), rare functionally disruptive missense mutations and splice-site acceptor/donor mutations (between 40-60 and ≤ 2 respectively, internal data from 9 whole-exome sequenced individuals – unpublished data). Due to the elevated probability of an allele being homozygous in a consanguineous individual, it is likely that at least one gene will be completely dysfunctional/inactivated by these rare mutations (Figure 7.2).

Knockout studies in model organisms are well established and have hugely facilitated our understanding of our genome and the biological pathways which connect some of these genes. However, where not backed up by human observational studies, animal knock-outs can be misleading as the underlying mechanism may be different in the model organism or the gene may have a different (or other acquired) function(s). Also some human genes lack homologues in the commonly analysed model organisms (some may even have no homologues, termed ‘orphan’ genes [461]) which is another limitation of these gene knockout studies [462]. Therefore candidate genes derived from model organism ‘knockouts’ cannot be directly translated to a human model until the same phenotype (and genotype) is also observed in humans.

However, sampling randomly from a consanguineous population will enable the identification of natural human knock-outs, enabling the identification of non-essential genes, genes which cause late onset disorders (e.g. highly penetrant

* More in individuals of African descent

mutations in certain genes causing certain cancers) and embryo loss* - alongside the Mendelian disease causal ones. In this sense, studies of consanguineous populations can be classified as examples of a '*quasi*-reverse genetics' study (QRG), with direction of study being 'genotype to phenotype'. To put simply, which genes have been completely inactivated in a consanguineous individual can be determined initially using WES (or WGS where feasible), then the short-term and long-term effects can be observed, if any (see Figure 7.1 for examples).

*Obtaining tissue from miscarriages will have ethical implications which is outside the scope of this thesis

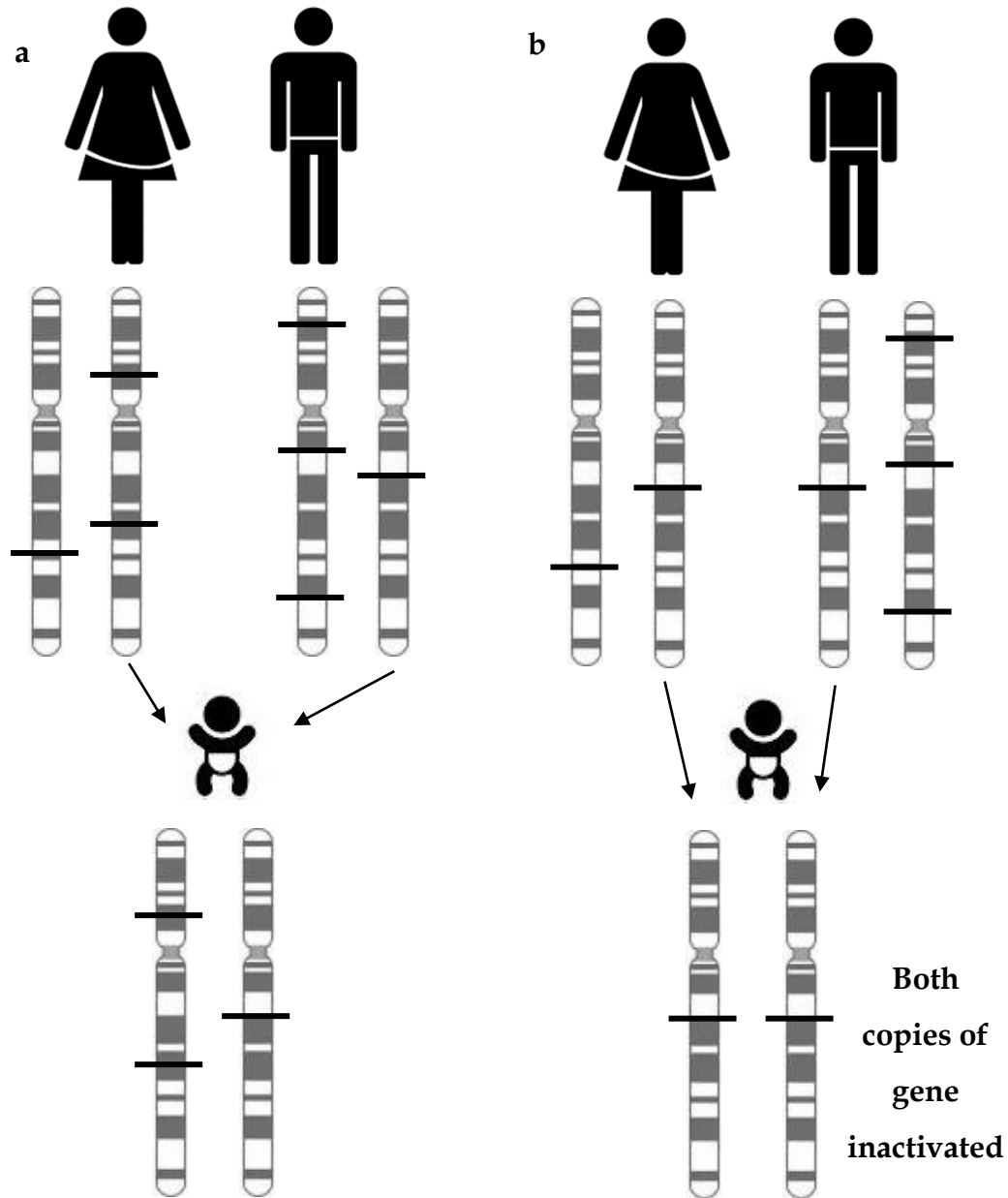


Figure 7.2 Example of difference between union of (a) unrelated (b) related individuals. Although everyone possesses LoF mutations within their genome, they are likely to be unique to their family. Therefore the offspring of unrelated individuals have an almost zero probability of being homozygous for these variants. Since related individuals will have a fairly recent common ancestor, their ancestors' LoF mutations will be passed on and there is on average 6.25% chance of these mutations to be in a homozygous (or more correctly, autozygous) state in the offspring of first cousins. Thick black lines represent LoF mutations. The figure has been simplified for clarity (e.g. does not include recombination events).

7.4.1. Overview of literature

Empirical studies suggest that next generation sequencing (NGS, introduced in section 1.6.2) of the whole genome of any individual (including healthy ones) will lead to the identification of many rare loss of function mutations but in a heterozygous state. In an outbred population (which are, as abovementioned, are the main focus of most large-scale genetic studies/consortia), the homozygous effects of these mutations will mostly remain unknown. There are instances where genes have been ‘fully’ (i.e. loss of function mutation in a homozygous state) inactivated but without major clinical consequences. One such example is analbuminaemia, a rare autosomal disorder where albumin production is halted [463]. Although albumin is the most abundant protein in the human plasma constituting 50% of the total plasma protein content (and represents over 90% of transcripts in hepatocytes – the predominant liver cell type) [464], patients where both copies of the albumin (i.e. *ALB*) gene is inactivated, the clinical symptoms are always remarkably mild thus diagnosis at infancy is difficult [465]. Using this example as a way forward, in order to understand fully the biological and clinical importance of the twenty one thousand or so human genes, their effects must be observed when fully inactivated – similar to reverse genetics studies carried out in model organisms. Finding these loss-of-function mutations in a homozygous state will only ever be feasible in consanguineous populations and/or collections.

Historically, the genomic load of “lethal” and/or detrimental disease causal mutations* has been estimated to be at least one (nearly two) by Muller in 1948, and then eight by Slatkin in 1954 [466, 467]. Combining birth defect and disease data obtained from various degrees of consanguineous offspring with a mathematically well-designed approach, Morton estimated a figure of 3 to 5 recessive mutations per gamete per generation which could be translated into 6 to 10 per zygote (even though Morton states a wider range of 6 to 15) [467]. With recent exome sequencing

* A distinction between the two has rarely been made in historical literature and cause confusion where specificity is required

projects, these estimations have been shown to be on the conservative side, with figures ranging from 10 to 20 rare (either absent or very rare in the dbSNP database) nonsense mutations per individual depending on ancestry (i.e. individuals with Yoruban/ African ancestry carrying more mutational load than a non-African individual's genome) [212]. This figure is considerably increased with the addition of gene inactivating indels (mostly frameshifting, figure thought to be between 8 and 17, again dependent on ancestry similar to the abovementioned example [212]), splice-site mutations and deleterious missense mutations. However as aforementioned, these will mostly be in a heterozygous state in the offspring of non-consanguineous unions (with $F \approx 0$). However due to increased autozygosity, there is a higher probability that an allele will be in a homozygous state in the offspring of consanguineous unions. For example 1 in 16 alleles are expected to be in a homozygous state in the offspring of first cousins; and this figure is doubled in the offspring of uncle-niece unions or double first cousins (i.e. 1 in 8 probability, see section 9.7 for depictions of consanguineous unions). Therefore, given a large enough consanguineous collection (details below), there would be potential to observe complete inactivation of most autosomal genes. This could then enable identification of a variety of gene inactivation effects from ones influencing foetal loss and ones contributing to the 3.5% excess (infant and pre-reproductive) mortality due to consanguinity (as discussed in section 1.3) [86, 468-470], to ones which have late-onset effects (e.g. the cases of Alzheimer's disease and Huntington's disease) where, especially in the latter, very little has been published connecting consanguinity with complex disorders and/or late-onset disorders [62, 86, 105, 471, 472]. Surprising results may also arise such that many genes could turn out to be dispensable (i.e. without major effects). As for genes on the sex chromosomes (mostly on the X chromosome) all gene inactivations should have been observed in males in outbred populations – since males are haploid (i.e. have one copy, thus are in hemizygous state) for both of them. However due to disease phenotype ascertainment, many late-onset and/or dispensable sex chromosome inactivations would have slipped under the radar and went unnoticed. This is reflected by the fact that out of the 1098 genes in the X chromosome, nearly half of the genes do not have

an assigned disease phenotype(s) [52, 473]; and many of the ones which do have been associated with intellectual disability (see Chapter 4 for further information) demonstrating the importance of X chromosomal genes in brain and mental function [474].

The following two sections were introduced and discussed in previous chapters (especially Chapter 1) but a short and relevant summary is included here for clarity of this chapter.

7.4.2. Effects of consanguinity on Mendelian disease

Very rare recessive mutations are predicted to be present in every population but since they rarely achieve homozygosity in outbreeding populations, they are mostly passed onto future generations silently. However, unions amongst relatives dramatically increase the probability of being homozygous at any genetic locus in the offspring (Figure 7.3). This is why very rare autosomal recessive disorders are predominantly observed in regions where there are high levels of endogamy or in families where the parents are closely related. Studying these populations will increase considerably the number of homozygous gene knockouts identified.

Erzurumluoglu *et al* have recently published a review on how best to identify highly-penetrant disease causal mutations which also provides an analysis schema to make the process more efficient and reliable [410]. The schema can also apply to monogenic forms of common-complex disorders (e.g. mutations in the leptin gene and obesity [475]).

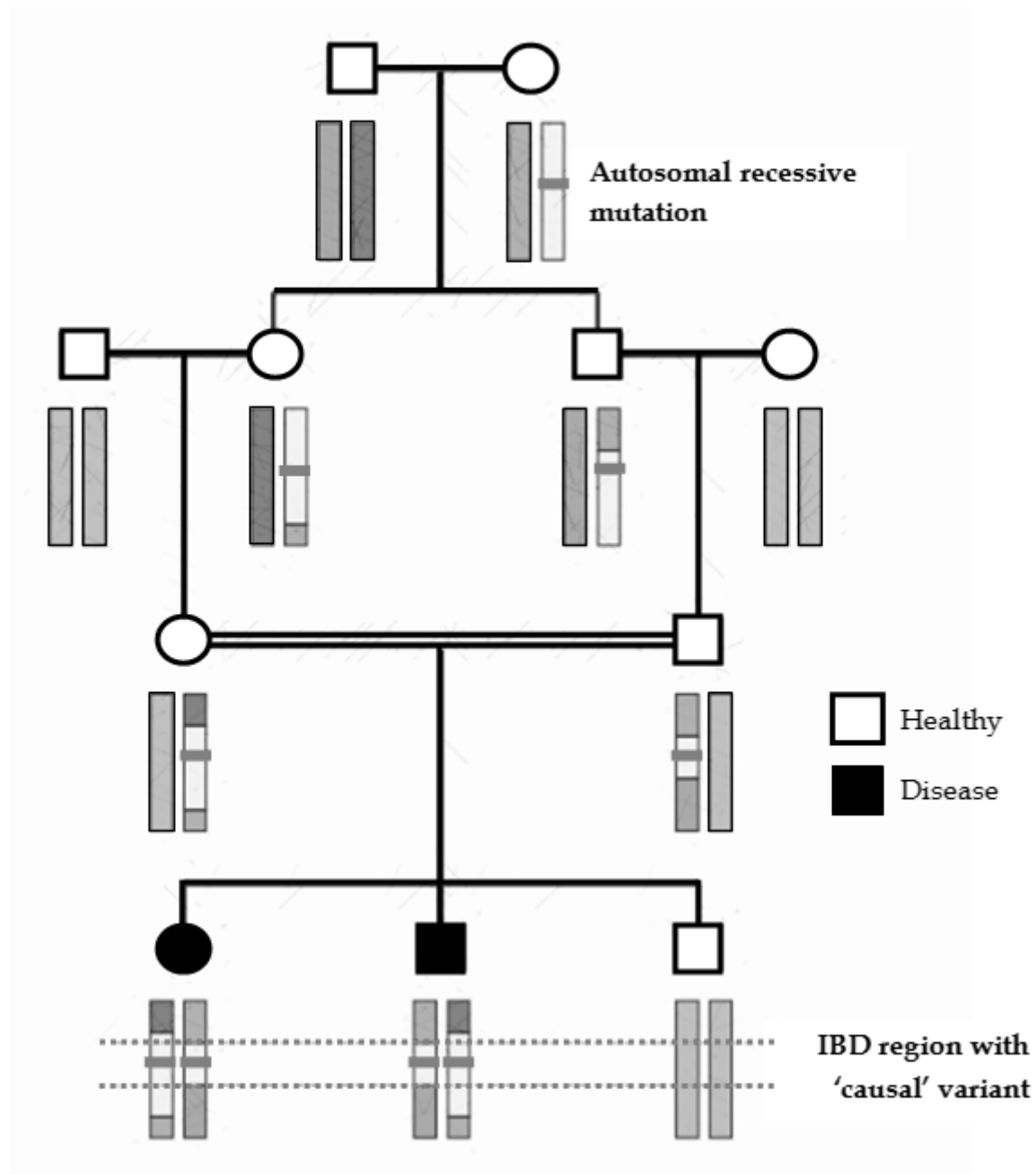


Figure 7.3 Consanguinity and increased homozygosity due to a recent common ancestor. A recessive mutation which is inherited within an inbreeding family has the potential to be passed down the generations and be inherited in a homozygous state (due to IBD) in the offspring of the grandchildren (first cousins). Assuming familial data is available, following the autozygous regions in the generations will enable researchers to pinpoint where the causal variant is; or LRoHs can be detected. IBD: Identical by descent.

7.4.3. Effects of consanguinity on common-complex diseases

The significance of human consanguinity on complex disorders *per se* is largely unknown; and the literature on the subject matter is inconclusive [86]. The role consanguinity plays on complex disorders is likely to vary depending on which model (e.g. infinitesimal model, rare allele model, broad sense heritability model – see reference [84] for details) explains the genetic basis (i.e. true underlying biology) of the complex disorder analysed. Consanguinity would be expected to have a greater influence on a complex disease if the rare variant model is to explain the aetiology of the disorder. A varying but lesser effect would be expected in the broad sense heritability model in accordance with the influence environmental factors (e.g. epigenetic factors, gene-environment interactions) have on the disorder analysed. With the infinitesimal model, one would predict consanguinity's effect *per se* on the disorder to be of a very small effect brought about only due to higher levels of homozygosity in consanguineous offspring (i.e. the effect of the minor allele is doubled in homozygotes compared to heterozygotes). This is because approx. 15/16 (93.75%) of the genome remains relatively 'outbred' even in offspring of first cousins.

In order to reliably deduce the role consanguinity plays in complex disorders, many environmental factors have to be considered as well as genetics and health related factors [105]. Consanguinity's effects on complex disorders *per se* cannot be reliably analysed through simple consanguineous vs non-consanguineous population comparisons (as has been done previously) – and many factors need to be controlled for (see Figure 1.10 in Chapter 1 for details) [86].

7.5. Quasi reverse genetics studies in humans?

As previously mentioned, gene knockout studies in model organisms have hugely facilitated our understanding of our genome and the biological pathways involved. However, where not backed up by human observational studies, they can always

mislead us as the underlying mechanism may be different in the model organism or the gene may have a different function. Also some human genes lack homologues in the commonly analysed model organisms which is another limitation of these gene knockout studies [462]. Therefore observing the homozygous effects of inactivating alleles in humans is crucial if pertinent progress is to be made in human genetics. For this to happen, consanguineous populations must be paid more attention than they are getting at present [86]. Due to significantly increased autozygosity in the offspring of consanguineous unions (meaning first cousins and closer in this case), they represent a considerable increase in the probability of observing the homozygous effects of inactivating alleles (see section 1.3.4 for more information on autozygosity), and thus increased probability of observing the effects of completely knocking out a gene in humans. Because there are many loss of function mutations in all individuals' genomes, analysis of consanguineous populations can be viewed as a QRG study with the direction of study being genotype to phenotype. Researchers can find out which genes have been completely inactivated by analysing the exome of the offspring of consanguineous unions and then observe what effect it will have had or will have in the future. Also by including individuals who do not display already identified clinical genetic disorders of childhood (for which there are more than three thousand genes identified [52, 476]), the probability of observing informative inactivations in the remaining (~17000) genes will be increased.

7.5.1. Frequency of natural gene knock-outs

To understand clearly what analysing consanguineous collections offer for human genetics, a comparison between an outbred and a consanguineous collection must be made (Tables 7.1 and 7.2). Consider a hypothetical outbred population in H-W equilibrium for a wild-type (and common) allele of frequency p and an inactivating allele of frequency q (i.e. the rare allele), where $p + q = 1$. Homozygotes for the rare allele will be found at frequency q^2 . However, in a consanguineous collection with a certain $\bar{F}$ (average inbreeding coefficient), an allele with frequency q will be expected to be in a homozygous state at a frequency of approx. $\bar{F} \times q$ (i.e. the overall likelihood of autozygosity for any given allele multiplied by the frequency of the

rare allele in the population – see row 1 of Tables 7.1 and 7.2 for a more accurate calculation).

Tables 7.1 and 7.2 illustrate the differences in homozygote frequencies between outbred and consanguineous populations for alleles with a range of different frequencies. It is clear that there is a higher probability of observing a homozygote for a rare inactivating variant in a consanguineous collection (see column 6 of Tables 7.1 and 7.2). In contrast there is a negligible chance, even with a large sample size, of observing a homozygote for a rare allele in a randomly breeding population. See Figures 7.4 and 7.5 for a comparison of alleles with MAF of 0.1 and 0.001 in consanguineous populations. For example, for a disorder such as homozygous Familial hypercholesterolaemia with a global prevalence of 1 in a million [477], according to the H-W equation one would estimate the frequency of the causal allele (i.e. q) to be 1 in a thousand (such as in row 5 of Tables 7.1 and 7.2 below). However in a consanguineous population, this figure (of 1 in a million) will be inflated approximately 60 fold to around 1 in 16000 (~120 fold to 1 in 8000 in a collection of offspring of uncle-niece unions and/or double first cousins).

Heterozygotes for alleles with MAF between 0.05 and 0.001 are expected to be found in large sequencing projects such as the 1000GP and some of these alleles may be worth searching for (i.e. ones suspected to be inactivating a gene without a disease assigned) in a consanguineous population for homozygotes through direct genotyping (compare rows 4 and 5 in Tables 7.1 and 7.2). However the inefficiency of trying to explore the homozygous effects of rare gene inactivations in an outbred population is epitomised in rows 4 to 8 in Tables 7.1 and 7.2 (also see column 6 in Tables 7.1 and 7.2). Also consanguineous populations (such as the ones mentioned in section 1.3.5) are not well represented in current sequencing projects (e.g. 1000GP) and are biased towards Western and/or Far Eastern countries; and this will cause missing out on ‘unique’ and/or clinically relevant alleles present in these (consanguineous) populations, thus carrying out whole exome sequencing (WES, see section 1.5.2 for more information) of consanguineous populations will allow identifying these unique alleles – and better – in a homozygous state; and is the most

suitable for achieving the aims of such a study. A DNA bank (with WES data) of 10000 participants (who are offspring of consanguineous unions equal to or closer than first cousins) would represent a resource of thousands of (different combinations of) gene inactivations in unrelated individuals (Equation 7.2).

$$\text{Total } G_{inactive} = \bar{G} \times \bar{F} \times N$$

Equation 7.2 Calculating total number gene inactivations (i.e. Total $G_{inactive}$) in a consanguineous collection. $\bar{G}$: Average number of genes inactivated in individuals within a certain population, $\bar{F}$: Average inbreeding coefficient of the database, N: Number of participants.

Using Equation 4 above, one can make inferences about how many gene inactivations are to be expected from their respective DNA bank of consanguineous offspring. For example, one would expect between 18 to 37 gene inactivations in any individual depending on their ancestry (adding together the figures of 10-20 for rare stop gains and 8-17 frameshifting indels from Ng *et al* [212]). This would then be multiplied by the probability that any allele will be autozygous in the dataset which will be 6.25% (i.e. 1/16) for a collection comprising mostly of offspring of first cousins (and 12.5% for a collection of offspring of uncle-niece unions and/or double first cousins) and the number of participating individuals, which will be arbitrarily chosen to be ten thousand. Thus one would expect between 11 thousand and twenty four thousand (11250 to 23125 to be more exact in this example) complete gene inactivations caused by rare mutations in a collection consisting entirely of offspring of first cousins. This notable figure will be boosted with the addition of offspring of uncle-niece unions and double first cousins which will increase the average inbreeding coefficient, while structural variation, LoF missense and splice-site mutations will add considerably to the number of completely dysfunctional genes (MacArthur *et al* predict this figure to be 100 LoF variants in healthy human genomes [97, 478]). Furthermore, homozygous stop gains which do not cause NMD in clinically unaffected individuals can indicate exons which are not essential for gene function; and vice versa, can point to regions which are essential for development in clinically affected individuals. No matter what the addition from the other abovementioned sources will be, bearing in mind that there are around 21

thousand genes in the human genome [23], the significance of the figure calculated above is highly notable.

My (and Dr. Hashem Shihab at the University of Bristol's) base-by-base permutation analysis estimates that there are approx. 4.5 million potential stop-gains, approx. 78 million missense mutations (with over 30 million predicted to be deleterious by SIFT and Polyphen-2, and over 10 million by FATHMM) and approx. 0.5 million stop-losses; and presumably thousands of essential splice-site donor/acceptor variants to be observed in the human genome (see Table 7.3 for details) [242-244, 479]; and observing a sufficient number of them (i.e. at least one per each gene) in a homozygous state can only be feasible in consanguineous collections (see Tables 7.1 and 7.2).

1. Row	2. MAF (q)	3. Heterozygote frequency in outbreeding population ($2pq$)	4. Homozygote frequency in outbreeding population (q^2)	5. Frequency of homozygotes (of q) in First cousins' offspring ($q^2 + (1-q)qF$)	6. Relative odds of finding homozygotes ($((1+F(1-q))/q)$)
1	0.1	0.18	0.01 (1/100)	0.015625	x1.6
2	0.0316	0.0432	~0.001	0.00291259	x2.9
3	0.01	0.018	0.0001 (1 in 10000)	~0.000725	x7.2
4	0.00316	0.00432	~0.00001	~0.0002075	x20.7
5	0.001	0.0018	0.000001 (1 in a million)	~0.0000635	x63.5
6	0.000316	0.000432	~0.0000001	~0.00001985	x198.5
7	0.0001	0.00018	0.00000001	~0.00000626	x626
8	0.0000316	0.0000432	~0.000000001	~0.00000198	x1978

Table 7.1 A comparison between collections of outbred offspring and collections of offspring of first cousins. Offspring of first cousins are expected to have an F value of 0.0625, whereas the expected F value for the offspring of outbred individuals is (very near) zero. MAF: Minor Allele Frequency.

1. Row	2. MAF (q)	3. Heterozygote frequency in outbreeding population ($2pq$)	4. Homozygote frequency in outbreeding population (q^2)	5. Frequency of homozygotes (of q) in offspring of uncle-niece unions ($q^2 + (1-q)qF$)	6. Relative odds of finding homozygotes ($((1+F(1-q))/q)$)
1	0.1	0.18	0.01 (1/100)	0.02125	x2.1
2	0.0316	0.0432	~0.001	0.00482518	x4.8
3	0.01	0.018	0.0001 (1 in 10000)	0.0013375	x13.4
4	0.00316	0.00432	~0.00001	0.0004049605	x40.5
5	0.001	0.0018	0.000001 (1 in a million)	~0.000126	x126
6	0.000316	0.000432	~0.0000001	~0.0000396	x396
7	0.0001	0.00018	0.00000001	~0.00001251	x1251
8	0.0000316	0.0000432	~0.000000001	~0.000003951	x3951

Table 7.2 A comparison between collections of outbred offspring and collections of offspring of uncle-niece unions (or double first cousins).

Offspring of first cousins are expected to have an F value of 0.125, whereas the expected F value for the offspring of outbred individuals is (very near) zero. MAF: Minor Allele Frequency.

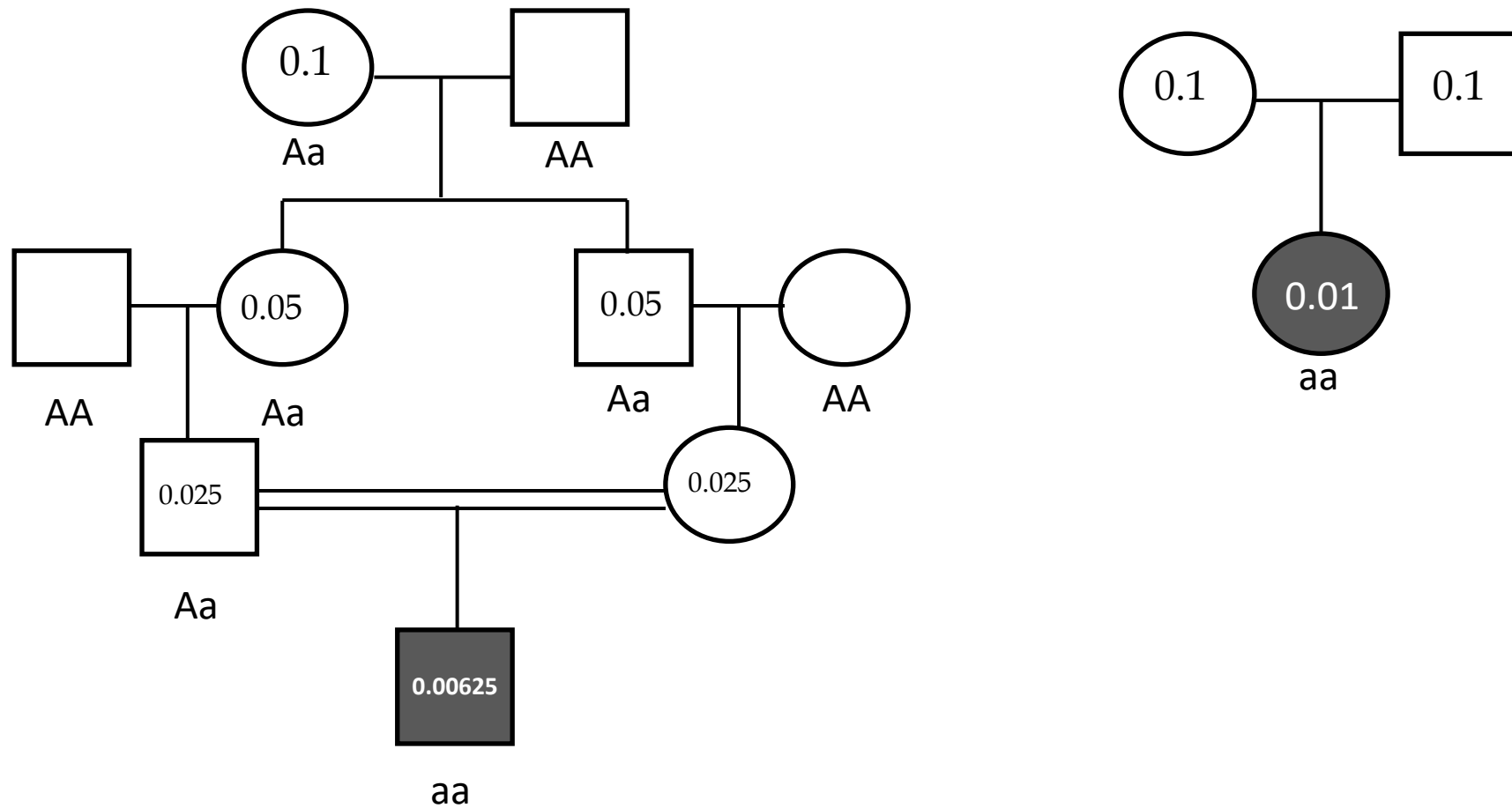


Figure 7.4 Comparison between offspring of outbred individuals and first cousins using the example of an allele for which $q = 0.1$ (frequency of 1 in ten in a population) and there are three unrelated homozygotes (i.e. AA) who marry into the family. In this scenario, when an allele is very common in a population (e.g. 0.1 as in this case), it is much more likely to find a homozygote in the offspring of outbred individuals compared to the offspring of first cousins. This is because of the lower probability of the LoF mutation travelling down the generations to meet its counterpart in her great-grandchildren. Due to purifying selection, we would not expect mutations causing LoF to be close to this frequency (see Figure 7.5). A: wild-type allele. a: LoF allele.

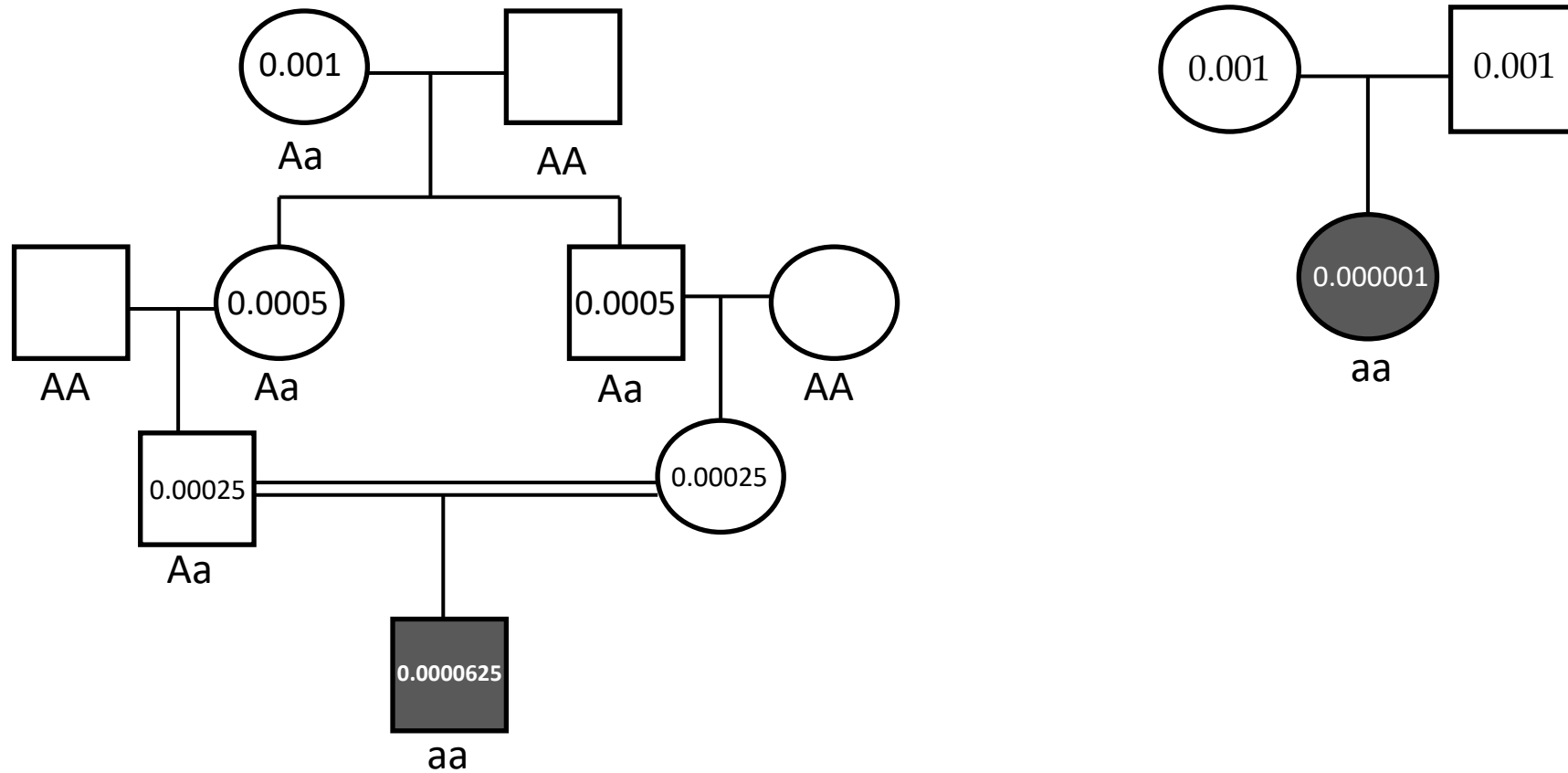


Figure 7.5 Comparison between offspring of outbred individuals and first cousins using the example of an allele for which $q = 0.001$ (frequency of 1 in thousand in a population). The true effects of consanguinity is seen in this example as there is 62.5 fold increased probability of observing a homozygote in the offspring first cousins compared to offspring of outbred individuals – even if the LoF mutation has to travel down three generations to meet its counterpart. A: wild-type allele. a: LoF allele.

Chromosome No	Potential stop-gains	Potential missense mutations	Potential missense variants predicted 'deleterious' by SIFT	Potential missense variants predicted 'deleterious' by Polyphen-2	Potential missense variants predicted 'deleterious' by FATHMM	Potential stop-losses
1	487480	8467065	3347260	3854115	1221832	62572
2	363376	6108312	2563015	2782135	995554	44020
3	283426	4849265	1874572	2198890	767750	42152
4	205941	3388227	1297317	1485985	473172	33151
5	219006	3738180	1553672	1759632	504748	25432
6	244104	4151134	1692057	1926158	572431	28653
7	220408	3847838	1564136	1695240	614793	28698
8	168650	2881145	1151638	1265275	418389	26880
9	189793	3370973	1333376	1524419	474573	24432
10	191621	3289834	1323421	1488356	468687	24363
11	262358	4850828	2001340	2210778	697377	33717
12	253130	4361643	1745694	2001575	757405	31582
13	87557	1448206	549045	679773	275827	9120
14	155852	2672365	1020606	1182304	342163	27263
15	155001	2686408	1169872	1298185	452945	8352
16	174677	3350511	1406819	1588250	512379	12994
17	242866	4510502	2010220	2158614	829781	19853
18	72225	1223623	504444	597802	164751	4140
19	274257	5160678	2450500	2416241	699383	19202
20	107404	1984320	811427	905090	298709	14196
21	49159	841258	331579	373539	121269	8331
22	92490	1785378	729532	802734	262976	11603
Total	4,500,781	78,967,693	32,431,542	36,195,090	11,926,894	540,706

Table 7.3 Potential number of LoF mutations in the human genome. To calculate the potential number of missense, stop gain and stop loss mutations in the human genome, we downloaded the dbNSFP database (release 2.6) and parsed the functional annotations for all potential non-synonymous single nucleotide variants. Predictions for missense variants were obtained from SIFT, Polyphen-2 and FATHMM.

7.5.2. Suitable Sample Populations

Isolated populations (e.g. living in islands with lack of long term genetic admixture) and/or endogamous populations which arose from small founder populations (e.g. Amish, Hutterites, Druze) have been under scrutiny for a long time. If a study similar to the one mentioned in this chapter was to be carried out in these populations (i.e. WES of offspring of consanguineous unions between kin related closer than or equal to first cousins), one would expect to see similar genes being inactivated due to a relatively small gene pool (e.g. the case of the 'Finnish disease heritage' in the Finnish population which is considered to be 'homogeneous isolate' [333-335]). This is because there are bound to be founder mutations in a lot of seemingly unrelated individuals due to reduced diversity and homogeneity which will decrease the efficiency of the study and may lead to waste of a considerable amount of time, funding and effort [480]. Thus the chosen populations must be chosen in an astute way so that they have a large population with a considerable amount of recent migration (thus with a rich gene pool), but also where consanguinity is prevalent – and preferably where it has a long history. This will mostly eliminate the problem of unrelated individuals turning out to be cryptically related* after WES has been carried out on their genomes [481]. Of course, the populations also should not have been analysed (at large-scale) genetically previously. A few suitable populations are presented and discussed below:

City of Riyadh

Located at the centre of the Arabian Peninsula and being the capital as well as the largest city of the Kingdom of Saudi Arabia (KSA), Riyadh has an ever increasing population size – with current estimates reporting an urban population figure of over 4 million (Figure 7.6). However early in the twentieth century, the city's population was a mere twenty seven thousand [482]. This dramatic increase in population is due to three very influential factors: Large family sizes (i.e. average

* covariance between seemingly unrelated individuals because of their unknown relatedness

size is above 6 for Saudi families and approximately 5 for non-Saudi families), rapid economic growth and immigration (e.g. Asians such as Pakistani and Indian, and Arabs from Yemen and Egypt). Tens of thousands of (mostly non-Saudi) rural dwellers still continue to migrate to the city of Riyadh each year [482]. This influx of families from around the Arabian Peninsula translates into a very rich gene pool, important for the abovementioned reasons. Furthermore Riyadh could be called a 'mecca' for consanguinity with over 50% of total marriages being consanguineous, including thirty to forty percent first cousin marriage rate ($\alpha = 0.023$) [86, 146, 483]. Within-family marriages are a deep rooted tradition thus observed autozygosity may turn out to be higher compared to standard estimates (meaning higher probability for an allele to be homozygous), which is another advantage of carrying genetic analyses in Riyadh. The quality of life is high in Riyadh (with access to advanced medical care and good communication services) and the country is relatively stable politically, economically and geographically compared other countries in the region [482]. The King Saud University which is the leading university in the Arab world (according to the QS World University Rankings 2013, available at <http://www.topuniversities.com/university-rankings/world-university-rankings>) and the King Faisal Specialist Hospital and Research Centre with its established centre for consanguinity studies is also located in Riyadh, which is important for possible collaboration. The initiation of the Saudi Human Genome Project (<http://shgp.kacst.edu.sa/site/>) is also an important platform for collaboration opportunities.

For a more comprehensive review on the genetic studies carried out in the KSA and the infrastructure that is available, see reference [484].



Figure 7.6 Location of Riyadh in the Arabian Peninsula and KSA. Image reproduced under the Wikimedia Commons Licence, source URL: <http://en.wikipedia.org/wiki/Riyadh>.

Cities of Andhra Pradesh and Karnataka

Located in the South Eastern part of India (Figure 7.7), Andhra Pradesh has a population size of over eighty four million [485]. The city has a highly diverse population with many languages spoken from Telugu (the language of Andhra people) and Urdu (language of Pakistanis) to Hindi (language of modern day Indians) and Tamil (language of Dravidian Indians). Next to Andhra Pradesh is Karnataka with also a diverse population of size of over sixty one million. The Hindu segment Andhra Pradesh and Karnataka population show a remarkable contrast in the rate of consanguinity compared to other parts of India, where the overall rates have been low and/or diminishing in the latter (see Figure 1.13) [86, 485]. Especially the rates of uncle-niece marriages reach as high as twenty percent of total Hindu marriages in Karnataka and approximately 5% in Andhra Pradesh – this

is an important feature for genetic studies as their offspring are expected to be homozygous (i.e. autozygous) for 12.5% of their genome [133, 486]. Also health facilities have improved a lot in both cities due to continuous government funding [485]. Average family sizes are also higher compared to European families (2.6 in India according to the Population reference bureau, see <http://www.prb.org/>). Consanguinity rates in Andhra Pradesh, Karnataka and Tamil Nadu are 30.8% ($\alpha=0.0212$), 29.7% ($\alpha=0.018$) and 38.2% ($\alpha=0.026$) respectively [170].

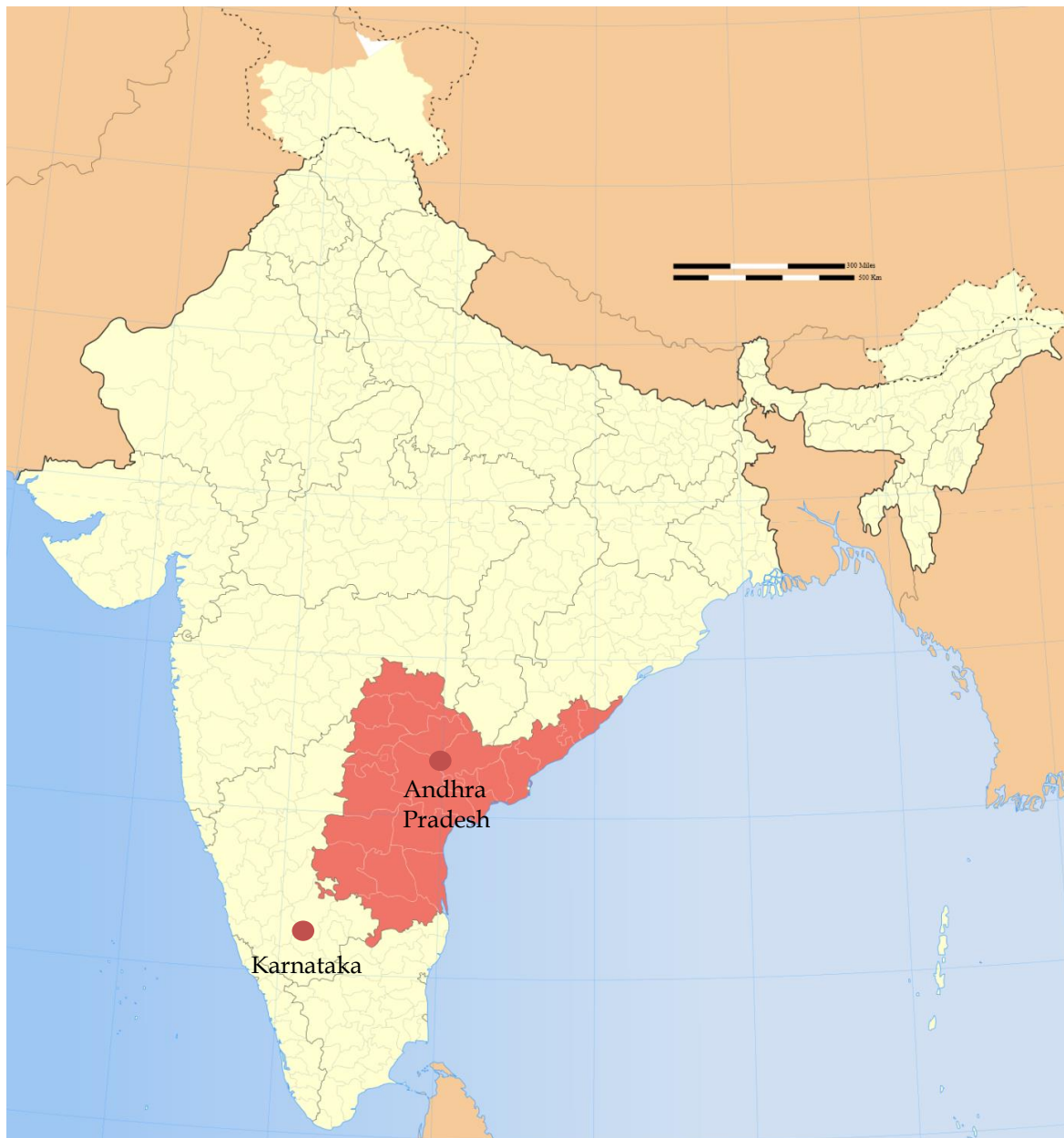


Figure 7.7 Locations of Andhra Pradesh (in red) and Karnataka in India. Image (adapted and) reproduced under the Wikimedia Commons Licence, source URL: http://en.wikipedia.org/wiki/Andhra_Pradesh.

Pakistan

Pakistan as a whole has a very high rate of consanguinity (over 40%) [105] . However at current, it may not be feasible for QRG studies at present, not due to genetic and/or clinical factors but for political reasons (e.g. periods of military rule, conflicts with India, corruption) [487]. With a population of over 180 million individuals and judging by the amount of infant/childhood deaths and autosomal recessive disorders in the Pakistani population living in the UK (especially the city of Bradford) [488], large scale studies carried out in Pakistan are destined to uncover many gene inactivations. Average family sizes are much higher compared to European families (3.6 in Pakistan according to the Population reference bureau).

Others

It may not be feasible to carry out large scale studies in many cities at once thus small scale collaborations can be initiated in other populations where consanguinity rates are high such as in the Bedouin tribes/communities of the Arabian Peninsula (e.g. in Oman consanguinity rates reach as high as fifty percent [489]), certain populations in Bangladesh and in (mostly unexplored) tribes of Africa which is bound to increase the mutational spectrum of QRG studies (see Figure 12).

Consanguinity rates in other countries:

Turkey (mostly in Eastern Turkey): 20.1% ($\alpha = 0.011$) [137]

Sudan: 52% ($\alpha = 0.0302$) [490]

Jordan: 58.1% ($\alpha = 0.036$ – assuming all first cousin marriage) [491]

UAE: 50.5% ($\alpha = 0.0222$) [182]

7.6. Methods

7.6.1. Ethics

All data (genotypic and phenotypic) will be anonymised (i.e. assigned an ID number) thus researchers will not know which individuals' data they are analysing.

Informed consent and approvals will be obtained where ever blood samples are taken and analyses are being carried out.

7.6.2. Creating a DNA Bank

The price of WES is a lot cheaper than WGS at present; for example BGI-Tech was offering WES for \$899 per sample (see <http://bgiamericas.com/service-solutions/genomics/exome-target-regions/>) and this figure is expected to be lowered when thousands of samples are involved. Only eligibility criteria will be that all participants must be either offspring of first cousins, double first cousins or uncle-niece unions. DNA will be extracted from obtained blood and sent for WES. As for phenotyping, basic details of contact details, individual health and family health will be recorded at regular intervals throughout life. Once WES data is received, phenotype and genotype data can be merged (keeping a copy of the original data untouched) using software such as Plink [294], Stata [492] or VCFtools [251].

7.6.3. Identification of Autozygous regions and Gene inactivating variants

WES data can be analysed for LRoH directly using software such as AutoZplotter, Plink, AgilentVariantMapper and AgileGenotyper [294, 296, 410]. To use the former two tools, sequencing data should be reformatted in VCF file and Ped/Map format respectively. Once autozygous regions have been identified, one could identify which genes contain 'predicted' deleterious missense mutations, splice-site mutations and/or nonsense mutations (see Φ mutations in Chapter 3 for details). 'Deleteriousness' can be predicted by using software such as FATHMM [242], SIFT [243] and Polyphen-2 [244] (see section 2.4.3 for more detail on mutation effect predictors). Then these derived datasets can be merged to see which one of these 'deleterious' mutations fall within the autozygous regions determined before. Once the prime candidates for complete inactivation have been identified, post genotyping (and functional analyses where feasible) will be carried out to make sure the

inactivations are real/present. The same procedure will be repeated for each individual and a set of 'inactivated genes' will be determined for each one of them.

7.6.4. Comparative Genomics

Once a gene is characterised as 'inactivated', if the gene has no disease and/or function assigned to it in humans, then deeper phenotyping could be carried out. Also a homologue in another model organism could be 'knocked out' to observe the phenotypic effects which will provide evidence on potential functions the gene carries and the pathways the gene's product (i.e. the protein) is involved in. For this, several model organism knockout consortia are available such as the International Mouse Phenotyping Consortium (generates knockout mice, website at <https://www.mousephenotype.org/>), Flybase (generates knockout *Drosophila melanogaster*, website at <http://flybase.org/>), Wormbase (a database on *C. elegans* and related nematodes' genomes, website at <http://www.wormbase.org/>) and Zfin (a database on Zebra fish genetics, website at <http://zfin.org/>).

7.7. Discussion

Since many Mendelian disorders are rare and are caused by autosomal recessive alleles, regions where consanguinity is high (see reference [86] for a World map of consanguinity) need to be paid more attention. However we have also pointed out that selecting only for disease cases ignores genes without clinical relevance, ones which have subtle cellular effects and/or which may contribute to late onset disorders. For this reason we recommend also sequencing/genotyping consanguineous individuals who do not show any clinical features early in life together with those who do. It may even turn out that they harbour previously identified disease causal mutations but do not show any clinical signs as they also simultaneously possess (highly-penetrant) protective variants (e.g. against highly-penetrant autosomal dominant mutations which interfere with other pathways). We have also suggested designing a SNP array to serve this purpose.

In this chapter, the framework of a genetic epidemiology project was described which resembles gene knockout studies carried out in model organisms – thus coined the name ‘*quasi* reverse genetics’ (QRG) studies carried out in humans. The primary aim of the study is to identify and characterise as many gene inactivations as possible in humans. The direction of study will be genotype to phenotype – just as in gene knockout studies carried out in model organisms where offspring of consanguineous unions are statistically very likely to have a few genes completely inactivated, then the effects of knocking out these genes can be observed either immediately (e.g. if causal of a Mendelian disease such as Cystic Fibrosis), later (e.g. if causal of late-onset disorder such as Haemochromatosis) or never (e.g. dispensable gene). These studies can also inform (molecular) researchers on how environmental factors affect certain individuals in a way that they could observe whether there are any individuals with their gene (or biological pathway) of interest inactivated; and then they could follow up with certain clinical, physiological and/or cellular phenotypic studies. Single (or a few) instances can be highly informative about modifiable outcomes which are generalizable to the whole population. A fitting example is if the causal role of LDL cholesterol in coronary heart disease (CHD) could be elucidated earlier through extra focus on familial hypercholesterolemia which is associated with elevated levels of low-density lipoprotein-cholesterol (LDL cholesterol) which translates into increased risk of coronary heart disease due to the causal role cholesterol plays in CHD, then problems that the Western world faces with CHD could be targeted decades earlier as there have been publications on the matter since the 80’s [493, 494]. The results obtained from a single (or a few) individuals can be as important at the population level as findings from studies such as Mendelian Randomisation (MR) studies [495] which can require tens of thousands of cases and controls with huge costs to the funding bodies.

Another advantage of carrying out large scale QRG studies is, alongside identifying new causal variants and novel gene functions, identifying how common these variants are. Inferences from the results (e.g. MAF) could affect policy makers’

decision to adopt a pre-marriage screening where there is not one, or improve current screening tests with the addition of these new variants/genes.

7.7.1. Way forward?

Using a brute force approach to whole-exome (or whole-genome) sequencing as many consanguineous offspring as possible is presumably not going to be cost efficient as WES is still prohibitively expensive for (very) large-scale sequencing studies, and as many of the offspring will not harbour any 'distinct' LoF variants in a homozygous state. There is also a lack of consensus as to what defines a 'LoF' variant. Mostly, rare coding mutations which pass a certain arbitrarily chosen threshold for conservation (or predicted 'deleterious' by a certain tool) are being clustered under the name 'LoF'. However, where these variants are not followed up by functional studies (e.g. gene expression studies), the evidence for the variant being 'LoF' is usually very low and unconvincing.

Therefore we propose that a SNP array containing probes for (i) all *possible* NMD causing stop gains and (ii) all other known LoF and/or disease causal mutations may be designed and applied to as many consanguineous offspring. Homozygous stop gains which are targets for NMD (e.g. in the 5' end of the gene transcript and >55bp remains in the penultimate exon [496]) are highly likely to be LoF variants. Searching for these variants in a cost-effective manner is bound to increase the number of homozygous 'knockouts' identified in consanguineous populations. Such a SNP array will be better designed with expertise from different areas within the genetics field (e.g. model organisms, public databases). Additionally, all possible mutations with a CADD score of over 50 (arbitrarily chosen here, representing top 0.001% of predicted deleterious variants) [269] and/or predicted deleterious by FATHMM-MKL (arbitrarily chosen here, score of ≥ 0.98) [270] can be added to the SNP array to validate these (and similar) tools' predictive power.

Comparing the SNP array proposed here to the traditional approach of using SNP arrays to identify the autozygome of an individual followed by sequencing of these regions, the former approach has several advantages. The SNP array will

additionally pick up variants which are homozygous via endogamy and chance, whilst the latter approach will only identify variants in the autozygome – excluding very short autozygous regions which are not identified; the SNP array will pick variants in these regions also. Furthermore, when carried out at a larger scale, identifying the autozygome for each individual and then designing primers to sequence these regions will become an unfeasible task. Such a SNP array is bound to serve the purposes of this type of study as the power to detect novel homozygous LoF mutations will be directly proportional to the sample size. Once the feasibility and the efficiency of the arrays/study are confirmed, similar studies can be carried out in isolated and/or endogamous populations searching for more novel LoF variants in a homozygous state.

Given the very low costs of SNP arrays compared to WES (or WGS), there is a bigger scope for identifying ‘true’ LoF variants with the former as the sample sizes will be much larger for the same prices. However, we must stress that we are not comparing WES/WGS with the SNP array approach proposed here *per se*, but rather we are comparing the two approaches in terms of characterising more novel and homozygous gene knockouts.

7.7.2. Addition to literature

The traditional approach to consanguineous populations is to ‘cherry pick’ families where a Mendelian disorder is segregating. Although this approach has yielded many disease causal loci, the effects of most genes in the genome are still not observed when both copies are inactivated. This can be due to the abovementioned phenotypic ascertainment of families which prevents the identification of homozygous knockouts of other genes as they do not cause a Mendelian disorder (especially during childhood). Randomly sampling from a consanguineous population is bound to increase our understanding of the human genome by enabling characterisation of novel gene functions.

Previous studies have attempted to use nullizygous CNVs and whole-exome sequencing to identify dispensable DNA and genes in the genome [497, 498]. These

studies have served as small-scale 'proof-of-concept' experiments (with the traditional inclination towards disease phenotypes and/or other distinct traits) and therefore have largely gone unnoticed, thus much larger studies with deep phenotyping are needed to understand the importance of consanguineous populations for human genetics. Very recently, two papers were made available in BioRxiv which carried out similar studies to the one proposed in this chapter [499, 500]. Although the studies should be commended for their potential contributions to the literature, the criteria used by the authors to define 'LoF' mutations is based on strong assumptions; and there is not much functional evidence provided by the authors that the variants identified do indeed cause loss of function of the respective genes. The SNP chip array we propose here will concentrate on (homozygous) stop gains which are very likely targets for NMD, and therefore are very strong candidates for causing loss of function of a gene – thus providing a more solid platform for characterising novel gene functions. With the addition of known disease causal variants to the same SNP array, there is also the possibility of identifying (highly-penetrant) protective variants (with regards to their respective diseases/traits).

In this chapter, I have also provided a theoretical framework for calculating expected number of genes with complete loss of function taking into account variants of many types (i.e. all single nucleotide variation and indels). As opposed to the traditional approaches, this chapter underlined the importance of studying consanguineous populations 'as a whole'. Additionally, pinpointing suitable populations represents reliable stepping stones for future direction of such analyses. The chosen populations - to be most effective - must have a rich gene-pool (due to mass migration and recent rapid population increase), while also being highly consanguineous (and/or endogamous). Riyadh's population is a perfect example of one.

7.7.3. Considerations

It is hard to put a price on the inferences that can be made from such a consanguineous collection as there is no benchmark to compare it against. However as mentioned above with the familial hypercholesterolemia and its connections with the role cholesterol plays in CHD, even a single instance of such a finding can lead to better understanding of a common problem and save the lives and/or increase life quality of many individuals.

The cost of such a large-scale WES project would be in the 7 figure sums (£500 sequencing cost x 10000 samples + additional costs) which is a lot less than the cost of most large epidemiological cohorts. Also the costs would mostly be accrued at the start of the project and would require a much smaller portion for running costs – unlike other epidemiological cohorts.

As for the SNP array study, the prices are bound to be much lower with similar sample sizes.

7.8. Conclusions

In this chapter I made several arguments for carrying out large scale sequencing studies in consanguineous populations as a whole by underlining the inferences that can be made from such a collection. I also presented several populations where carrying out such studies would be most effective.

CHAPTER 8. OVERALL DISCUSSION AND CONCLUSIONS

This thesis is founded on seemingly different studies (e.g. diseases, concepts*, data types); but they are all connected through consanguinity. On one side there are the family based Primary ciliary dyskinesia (PCD), Papillon-Lèfevre syndrome (PLS) and autosomal recessive intellectual disability (ARID) studies, on the other there is the population based theoretical study advocating large-scale sequencing projects in consanguineous populations as a whole. Additionally, an extensive literature review was carried out to specifically find out the effect of consanguinity *per se* on human disorders. Also with this thesis, a generic guide to identifying highly penetrant mutations from next-generation sequencing data was presented – which is applicable (and was mostly applied) to the analyses carried out in this thesis. An autozygosity plotter (named AutoZplotter, see Chapter 3) was also developed and published together with the review article (or chapter) to assist studies dealing with NGS data obtained from consanguineous individuals.

A family based study was carried out in a family with six ARID affected offspring. Autozygosity mapping (complemented with brute force SNP elimination) was applied to the SNP chip array data of the affected offspring and a 1.5Mb long region on chromosome 19 was identified. The next step would have been to amplify and sequence the region however the study was not followed up after realisation that Alazami *et al* had already published their findings which also included the family

* Empirical *and* theoretical studies present

described here [446]. Therefore our findings have served as a replication of their finding and a validation of their (and our) methods.

The effect of consanguinity *per se* was found to be small (if not negligible) on human disorders (derived from the literature). To elaborate, many studies have not been able to fully adjust for potential confounders such as population stratification (e.g. presence of autosomal recessive mutations, which inflate the effect of consanguinity), life style (e.g. food, smoking, alcohol intake) and environmental factors (e.g. conditions in hospitals, prevalence of viral or bacterial diseases). In fact, so far the only strong evidence for the effects of increased homozygosity *per se* is found in human height studies. Joshi *et al* reported that increased homozygosity was associated with a decrease in height*, the effect being equivalent to offspring of first cousins ($F=0.0625$) being 1.2 cm shorter [501]. They stated that the possible genetic mechanism behind this observation may be directional dominance†. As these studies are still in their early stages and the molecular mechanism behind these observations have not been determined, I will not speculate further.

As most of the family based studies in this thesis were carried out using whole-exome sequencing, they will be discussed in a separate section below.

8.1. Discussion on disease causal genes and WES

In this thesis, nine whole-exomes were sequenced and analysed. All WES data were found to be of high and/or sufficient quality (see section 4.6.1 for detailed discussion). The reliability of the WES data facilitated the identification of several variants known-to-be human PCD causal or which are good candidates as causal variants. Although WES does not capture all variation within a genome (Table 1.2), most of the studies carried out here resulted in either replications of previous

* I do not know how cognition and educational attainment can be associated to increased homozygosity, other than confounding effects within the data – most probably due to subtle population stratification

† Directional dominance occurs when dominance of variants for human height (or a certain trait) is (on average) biased in one direction

literature or discoveries of novel genes/variants worth reporting and/or following up with further studies – which justifies the funding, effort and time spent on these analyses.

Six separate family based studies were carried out with all families having at least one PCD affected offspring. One of the families also had PLS affected offspring. These studies have yielded a known variant in a known gene (i.e. p.R136* in *DNAAF3*), a novel variant in a known gene (i.e. p.G734fs in *HEATR2*) and novel variants in novel genes (i.e. p.E309* in *CCDC151*, p.M263T in *MNS1*, p.R263* in *DNALI1* and p.E328* in *LRRC48*). Also analysis of one of the family's WES data did not result in any potential candidates as a causal variant.

The variants p.G734fs in *HEATR2* and p.E328* in *LRRC48* were identified in the same individual, however the study could not be followed up as the collaborating clinician had lost contact with the family. It will be interesting to see in the long run whether *LRRC48* will turn out to be a human PCD causal gene. The variant p.M263T in *MNS1* also could not be followed for the same reasons.

The variant p.R263* in *DNALI1* remains an enigma as *DNALI1* would be considered a prime candidate for a novel human PCD causal gene, however the variant was found to be a heterozygote in one of the affected siblings which is not in accordance with the autosomal recessive mode of inheritance of PCD. However if p.R263* in *DNALI1* is causal (on its own), then it is likely that the heterozygote sibling may have been a mosaic in the lung cells (i.e. homozygote for variant) as DNA was extracted separately from blood *and* saliva and results were the same*. If there was additional time and funding, it would be a sensible next step to extract DNA from lung tissue to check for mosaicism. It will again be interesting to see in the long run whether *DNALI1* will turn out to be a human PCD causal gene. Another interesting finding in the same family was the unusual ultrastructure of the cilia – as both

* A mix-up between mother and the respective offspring's DNA samples was suspected, until second collection of buccal swabs

dynein arms were completely missing in addition to some of the subfiber Bs*. It may be that there is a second unidentified mutation which acted in an additive fashion to the *DNALI1* variants, causing this strange phenotype.

The human PCD causal variant identified in *CCDC151* was found to be rare in the local population. Therefore it would not be cost-wise to add to regular genetic diagnostic arrays in Riyadh (similar to neonatal heel prick tests in the UK). However in families which have a history of PCD, a custom SNP chip for PCD screening may be designed and probes for the p.E309* variant could and should be included given the highly penetrant nature of the variant. Our finding will be more applicable to genetic diagnostics (including preimplantation) and counselling at present, and then PCD curative studies in the long term as it will facilitate our understanding of the biological pathways involved in causing PCD. Another point of consideration is why our *CCDC151* mutant cilia displayed a difference in the ultrastructure in relation to the cilia reported by Hjeij *et al* [431]. The cilia we analysed resembled the results from animal knockout models (as discussed in 4.5.7) where both the IDA and ODA were absent in the cilia. However Hjeij *et al* reported that their mutant cilia was only missing its ODAs. The loss of ODAs is an expected phenotype as *CCDC151p* as localised in ODA and its function is ODA targeting and docking [409]. However our observation raises the possibility about whether there was another unidentified mutation in our patients which caused additional defects in the ultrastructure of their cilia.

The family with the *DNALI1* variant also harboured a PLS causal variant. Using the already available WES data we reasoned that there was high probability that the PLS causal variant would have been present in at least one of the two datasets. Our expectations were realised when a rare heterozygous variant in the form of p.G300D was present in the *CTSC* gene – the prime candidate gene for PLS. The variant was then followed up and we discovered that the variant was previously reported by Zhang *et al* [452]. This was concrete evidence for the causality of the variant. The PLS

* However 9+2 structure was retained

causal variant identified in *CTSC* was found to be rare in the local population and the factors abovementioned for the *CCDC151* variant also apply to this variant.

Identification of genes or variants responsible for disorders and/or health related traits facilitate the development of better genetic tests to identify individuals who are susceptible to a particular disease (mostly for common complex diseases) and/or identify ones who have a 'carrier' status (for Mendelian diseases); and will enable better molecular diagnosis (and prognosis) of patients [59]. Identifying new clinically relevant variants and/or genes will also make the results from prenatal testing more reliable (as more relevant tests can be carried out) and enable better informed decisions. Identification of a gene responsible for disease and/or a trait usually represents the first step taken towards understanding the physiological role the gene's product (i.e. protein) plays and further analyses can also elucidate the disease pathway which can serve as a platform for the development of preventive and therapeutic interventions.

8.2. Considerations and Thoughts on Consanguinity

Despite its importance to genetic research, consanguineous populations have not received the attention it deserves. This could be due to a variety of reasons including social, political and economic, not overlooking the fact that consanguinity happens to be prevalent in countries which seem to not have a long history of intensive research investment (Figure 1.13); and presumably it is much harder for outsiders to carry out genetic analyses. The relatively larger family sizes in some of the regions with high consanguinity levels could unearth the effects of autosomal recessive mutations, including very rare (or even unique) ones*. Many of these disorders have no cure and this raises awareness amongst many young consanguineous couples and they seek preconception genetic counselling in fear of what may lie ahead for

* As the probability of observing at least one homozygote child equals $1-(0.75)^n$, where n is the total number of children

themselves and their offspring [125]. They are seeking answers to questions such as “Will our children be physically or mentally abnormal?” and “How can we prevent/cure it?” [113] which require emotionally as well as scientifically sound answers.

There is an ever increasing awareness on prevention of congenital disorders in these regions and this is leading to an increasing number of couples to seek counselling on consanguinity and the effect it can have on their marriage/union. For this reason, preconception and/or premarital counselling should be part of the training health care providers receive – especially in highly consanguineous regions as inconsistencies in counselling for consanguineous communities have been reported in the past [125, 127]. Despite all the issues with pre and post diagnosis, carrier status detection and genetic counselling interventions have been started in a number of countries [114, 131, 150, 163, 168, 502]; and some have even started to bear fruit with reductions in prevalence of inherited disorders [115, 119, 141]. It is clear that not every consanguineous union should be treated the same way; and there has to be a distinction between families with known genetic/inherited disorders and ones without. Access to reliable health records with full pedigrees can only help this cause. A guideline on this matter has recently been proposed by Hamamy [503]. A midway should be found between the genetic risks and the positive psychological and social effects of consanguineous marriages; thus educating consanguineous populations in a comprehensive manner is only one part of the story. Additionally, genetic counsellors dealing with consanguineous families should also be educated and be made aware of the social side of the issue, which will enable both parties to understand each other and help take the best informed decision for the couples and their subsequent offspring. Rather than discourage consanguineous unions totally (which is bound to be ineffective), helping couples through all sorts of counselling should be the main objective [149].

The literature suggests that consanguinity levels can reach as high as 70% in certain populations. However, as a person who has lived in Eastern Turkey where

consanguinity levels are thought to be just over twenty percent [133], I am inclined to think that where reliable state records are not available, consanguinity rates may be over inflated in some of these regions as many individuals treat their far relatives (e.g. third cousins, $F = 0.0039$) and/or individuals with the same surname as 'close' thus may tend to report them as intra-familial marriages – since what is declared by the surveyed is usually not confirmed by empirical analyses (e.g. determining inbreeding coefficient via genome-wide SNP arrays). This is also thought to occur in studies carried out in Arab countries [113]. Therefore it is suggested that to best (report and) compare consanguinity rates between different populations, mean inbreeding coefficients (α) and the rate of union/marriage between first cousins should be measured and used [125].

To conclude this section, judging by the sociological literature, it seems like consanguinity unions (especially marriage between cousins) are here to stay; and for a long time. Although there are certain genetic risks to consanguineous unions (as discussed and presented in this thesis), individuals who engage in consanguineous unions are being criticised much more than they seem to deserve (with substantial criticism of migrant communities, especially in Western countries [105]); and the potential/known social and economic benefits (for these societies) are never discussed at the public level, with the very small (additional) proportion of affected families making the headlines. However this thesis is in no way advocating consanguineous unions, but rather trying to understand why these societies still engage in these types of unions – even though the genetic risks are well known.

Given the advancements in genetic diagnostics, screening for variants known to be disease causal should be encouraged. This will help couples 'at risk' to take more informed decisions; and rather than discourage consanguineous marriages totally (e.g. calls for banning first cousin marriages in the UK's Pakistani community, see reference [502] for an example) – which presumably is not going to work as it has deep roots within these communities [504], couples with risk of passing on lethal

and/or disease causal mutations (with or without cure/therapy) should be made fully aware of the probabilities and the consequences of their actions.

8.3. General additions to literature

During the analyses carried out in this thesis, several unplanned contributions were made to the literature. For example the term ‘ Φ mutations’ were introduced to the field via the paper where we published *CCDC151* as a novel human PCD causal gene (see ‘List of peer reviewed publications’ for details) [80]. This term was required as an umbrella term to collect all variants with ‘predicted high impact’ consequences*. This term will hopefully help those who work in similar projects where mutations with many different consequences are screened for candidacy of causality. Furthermore, the p.(E309*) variant was identified as ‘causal’ amongst thousands of variants even though only a single proband was present, showing clearly the importance of making full use of the bioinformatics tools and the public variant databases available.

An unexpected analysis was presented in Chapter 6 where we used the WES data of two siblings who were not affected with Papillon-Lèfevre syndrome (PLS) to identify the causal variant in two other PLS affected siblings for whom WES data was not available. Also, the term ‘proxy molecular diagnosis’ was introduced to the field as a result of this study (and the paper which was published consequently; see section 6.6 for discussion). This analysis also allowed us to discuss and debate the ethics of carrying out proxy molecular diagnoses in the same paper.

The term ‘*quasi* reverse genetics’ study was also introduced to the field of human genetics in Chapter 7 (and the paper which was published consequently [457]), making use of the fact that the proposed studies in consanguineous populations mimic the genotype-to-phenotype studies carried out in model organisms.

* Since ‘predicted high impact’ can be abbreviated as PHI, I initially used PHI to shorten the term in the paper. Then I realised that it sounded like the Greek letter ϕ (read ‘phi’), but decided to use the upper case version Φ - as PHI was also all uppercase.

8.4. Overall Conclusions and Future work

In this thesis I presented results from seven family studies where different data types (i.e. SNP chip array and WES) and/or different disorders (i.e. PCD, PLS and ARID) were present. In a chapter on its own, I argued for carrying out large-scale sequencing projects in consanguineous populations as a whole rather than the traditional 'disease ascertainment' paradigm – as the latter can only take us so far in facilitating our understanding of the human genome.

Several additions were made to the literature arising from these studies and they have also been discussed here (i.e. Chapter 8) and in their respective chapters.

As for future work, it is important to understand why the variants found here cause the phenotypes they caused (e.g. PCD, situs inversus, PLS, ultrastructural differences in cilia) and if further time and funding were available, I would liked to have carried out functional studies in model organisms to observe what happens to the protein within the cell (and pathway) at the molecular level to cause these abnormalities.

Finally, it is inevitable that WES will be replaced by WGS in the not-so-distant future (reason stated in section 1.6) and it will be interesting to see how this change will affect the way we analyse Mendelian and complex disorders – as it will bring its own computational, bioinformatics and statistical problems. But the WGS era will also be an exciting one as we will most likely make ground-breaking discoveries and find answers to the many enigmas which remain unsolved (e.g. missing heritability, novel gene functions, genetic interactions, pleiotropy, non-coding DNA)

CHAPTER 9. APPENDICES

9.1. General appendices

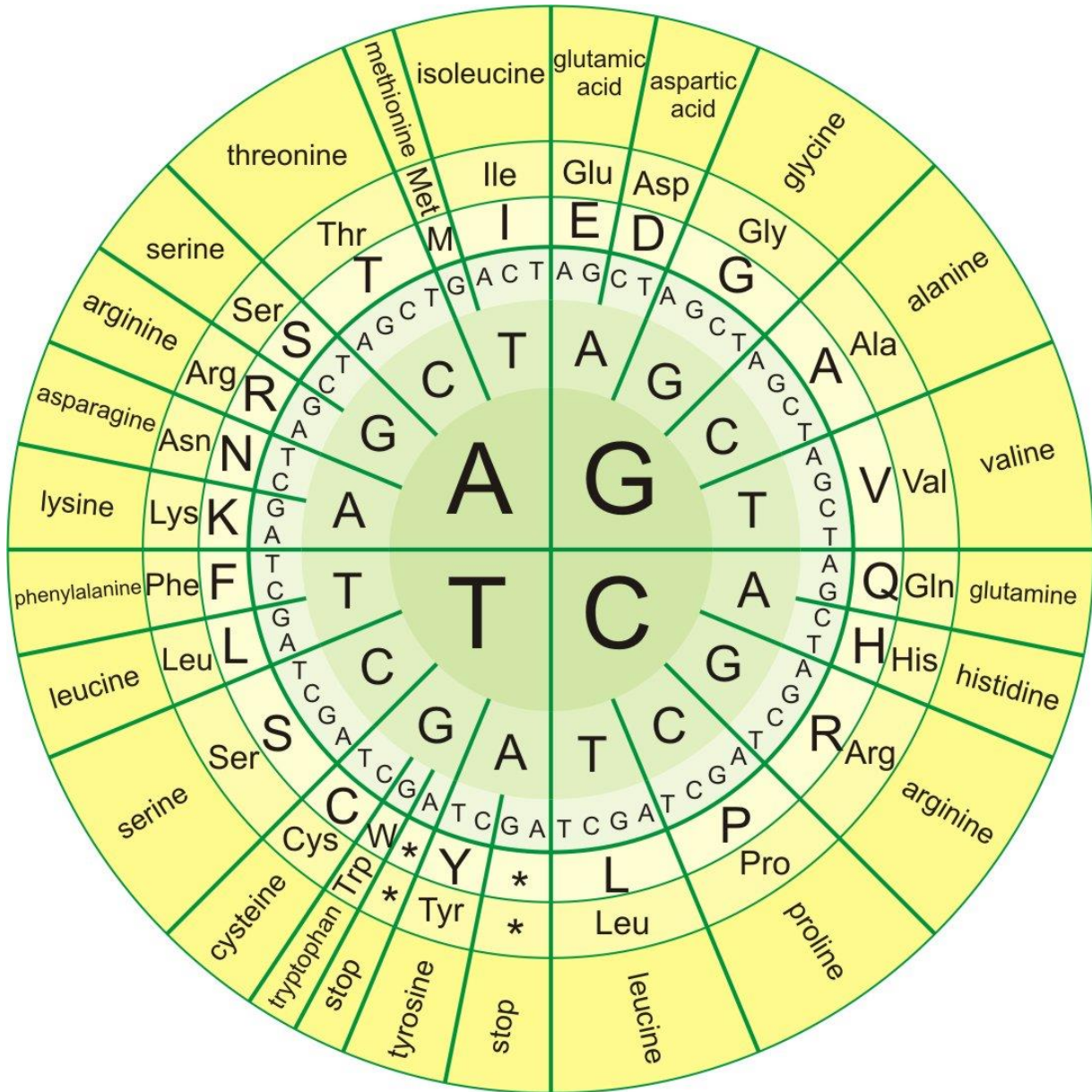


Figure 10.1 The Genetic Code: triplet of DNA bases codes for a corresponding amino acid or a start/stop codon. Image reproduced with permission, source URL: http://biol.lf1.cuni.cz/navody/molbiol1/genetic_code.jpg.

6/20/13

Medical Research Council



This is to certify that

Mesut Erzurumluoglu

completed the following e-learning with
an assessment (England, Wales and
Northern Ireland) score of

100%

Research and human tissue legislation

Overview of Human Tissue Act and Human
Tissue (Scotland) Act

When the Acts apply

What constitutes best practice

Top tips to support compliance

Where to find help

June 20, 2013

MRC Regulatory Support Centre

www.rsclearn.mrc.ac.uk/MRC_HumanTissueAct/data/swf/quizmaker_401/quiz_content/report.html

1/1

Figure 10.2 Certificate confirming HTA training received. Printed or pdf version is available upon request.

9.2. Baha'ism's view on consanguinity

National Spiritual Assembly of the Bahá'ís of the United Kingdom

27 Rutland Gate, London SW7 1PD

Tel: 020-7584-2566 Fax: 020-7584 9402

E-mail: nsa@bahai.org.uk

When replying please
refer to: GC-28748
28 November 2014
6 Qawl 171 B.E

Mr. Mesut Erzurumluoglu
epmmee@bristol.ac.uk

Dear Mesut,

Further to your email message of 24 October 2014, we have now looked into your very interesting question relating to consanguinity in the Bahá'í Faith.

Bahá'u'lláh, the Founder of the Bahá'í Faith, stated "It is forbidden you to wed your father's wives". This prohibition equally applies to marrying one's stepfather. Bahá'u'lláh also stated that further questions concerning the legitimacy or otherwise of marrying one's relatives would rest with the international governing body of the Bahá'í Faith, the Universal House of Justice.

Thus, subsequent to this, the House of Justice decreed the following law that is binding upon all Bahá'ís:

It is not permissible for a Bahá'í to marry his or her mother or father and their siblings and forebears, brothers or sisters and their descendents, or sons or daughters and their descendents. It is also not permissible to marry some corresponding categories of relations formed by bonds of marriage—such as the stepmother, stepfather, stepdaughter, or stepson, or the daughter-in-law, son-in-law, mother-in-law, or father-in-law—or similar categories of relations formed by legal and social bonds that create a shared family life, for example through adoption. Beyond these prohibitions, Bahá'ís should ensure they do not contract a marriage that would violate the customs or laws of the country in which they reside.

The House of Justice has refrained, at this stage, from defining other categories of relations with whom marriage is prohibited. It has stated however that it is important for Bahá'ís everywhere to be mindful of the following guidance in the Bahá'í Writings:

In marriage the more distant the blood-relationship the better, for such distance in family ties between husband and wife provideth the basis for the well-being of humanity and is conducive to fellowship among mankind.

We hope that the above information will be of use to you and wish you every success in your studies. May your research contribute to the advancement of knowledge and understanding of this subject, which no doubt has important implications for mankind.

Yours sincerely,



Patrick O'Mara,
Secretary

National Spiritual Assembly of the Bahá'ís of the United Kingdom
Registered in England – Company limited by guarantee No. 355737, Registered Charity (1967) 250851
Registered with the Office of the Scottish Charity Regulator (SC041673)
The Assembly also represents the Bahá'í communities of the Isle of Man and the Channel Islands

Figure 10.3 Letter received from the National Spiritual Assembly of the Bahá'ís of the United Kingdom detailing Baha'ism's view on consanguinity

9.3. Milestones in Genetics research

Genetic research, with the help of technological advancements and continually improved biochemical methods, has saved the lives and/or increase the quality of life of millions of people – whether it is through facilitating prevention via development of vaccines or by curing through the use of antibiotics, antiviral or antifungal drugs. It is impossible to say which breakthrough in genetics (and related fields) is more important as they are all complementary and required one another for each other's breakthroughs. Thus they should all be acknowledged on their own and not be compared with the others in this sense.

Modern genetics seems to have started when the results of the pea experiments of Mendel were published in 1866 – although the importance of his studies was not acknowledged in his life-time* [505]. After a decade of debate on where the 'genes' resided, Thomas Morgan confirmed that they resided on chromosomes in 1910 by proving that some characteristics of the *Drosophila* appear to be sex-linked – which marked another milestone in genetics. 16 years later Hermann Muller discovered that X-rays caused heritable diseases leading the way for studies on how this observation may be explained (i.e. they caused mutations in the genome). The year was 1944 when DNA was proven to be the actual hereditary material and not the proteins that reside on chromosomes (i.e. chromatin) also [18]. Probably the most famous of all discoveries which took place in the area of genetics happened in April 1953 when the correct structure of the DNA double helix was published by Watson and Crick [7]. Proving that the sequence length of nucleotides in DNA exactly matched a third of the sequence length of amino acids in protein (1964) and the first gene to be isolated (1969) were couple of the milestones which took place in the 1960s. The 70's were the beginning of the 'genetic engineering' era where the first gene to be synthesised from scratch (1970), first gene insertion experiments (from an African toad to a bacterium in 1973) and the first human gene to be cloned (insulin

* Lenay wrote a very nice paper on the re-discovery of Mendel's studies

gene in 1978) all occurred. However another advancement in the field of 'DNA sequencing' would take all the headlines at the end of the 70's when Maxam and Gilbert, and Sanger developed two sequencing techniques (see section 1.6); and the latter technique (with a few advancements) would later pave the way for every geneticists dream at the time to become a reality – The Human Genome Project. Significant advancements in the ethics of genetic engineering were also made during the mid-70s. The eighties would pick up from where the 70's advancements left and many advances in the genetic engineering area were made. The discovery of the polymerase chain reaction (PCR) in 1985 by Kary Mullis represents a true milestone [207]. Other milestones were the approval of insulin as the first genetically engineered drug (1982) and hepatitis B vaccine as the first genetically engineered vaccine for humans (1986), the development of first automated DNA sequencers (1986) and the location of a genetic marker for the Huntington disease (on chromosome 4, in 1983). The nineties started with the formal launch of the HGP (1990) and the discovery of the *BRCA1* gene (in 1991, which was shown to be located on chromosome 17 and was found to influence the predisposition of breast and ovarian cancer for certain individuals). The complete genomes of the *Haemophilus influenza* (Pfeiffer's bacillus, in 1995), *Saccharomyces cerevisiae* (Baker's yeast, in 1996) and *Caenorhabditis elegans* (roundworm, in 1998) were also sequenced in the 90s. The first microarray technologies were also developed (1995) in the 90s but the one milestone which attracted everyone's attention was the cloning of 'Dolly the Sheep' (1997). The second millennium would be named the 'post-genomic' era as the first drafts of the human genome were made available in June 2000 and later in Feb 2001. As all the excitement was directed towards the human genome, the publication of the *Drosophila melanogaster* (fruit fly, 2000) and *Mus musculus* (mouse, 2002) genome went unnoticed for many, although their impact on genetics cannot be overlooked as they served as 'model organisms' for years and still serving. 25th April 2003 marked the true beginnings of the post-genomic era as the mapping of the genes in the human genome were declared 'complete'. However the 'complete and annotated' sequence of the last chromosome (biggest of all, chromosome 1) of the human

genome was published in May 2006 [506]. The completion of the HGP set the stage for determining the function of the ~21000 genes that were identified in the HGP.

Incidental findings such as Walter Flemings observation in 1882 of ‘tiny threads’ (which are chromosomes) which appear to be dividing in the nuclei of salamander larvae and the discovery of DNA fingerprinting by Alec Jeffreys in 1985, stem cell research throughout the late 20th and early 21st century, not forgetting the development of large publicly and/or privately funded genome research centres (e.g. National Centre of Human Genome Research in 1989, Wellcome Trust Sanger Institute in 1993) have also speeded up advancements in genetic research and helped attract the attention of the lay media and the public towards genetic research. Genetics also has a few dark spots in its history as it has been proposed by some (especially Francis Galton) to be used to improve the human race through ‘eugenics’ – which ultimately ruined the lives of many children and led to the death of many individuals especially near and during the second World war.

Genome News Network have published a timeline for Genetics and Genomics at the http://www.genomenewsnetwork.org/resources/timeline/timeline_overview.php link which represents a valuable resource for understanding how genetic research came to the state it is today. Life Technologies have also made an “Illustrated DNA milestone” poster available to download and/or order (for free) at the <http://www.lifetechnologies.com/uk/en/home/brands/applied-biosystems/dna-anniversary/dna-milestone-poster.html> link.

9.4. Appendices for Chapter 3 – NGS data analysis guide

These commands are here to guide the user. However where complications arise, other options may have to be included thus requires reading documentation provided by the bioinformatics tools.

For users who are not familiar with UNIX commands and programming languages, the Galaxy server (<https://usegalaxy.org/>) provides a user-friendly interface by providing 'push-button' features for use for the NGS read QC, NGS read alignment, variant calling and variant annotation stages. Many other related UNIX-based features (e.g. Text manipulation, Filter and Sort) have also been adapted for use with the click of a button.

For more advanced pipelines and documentation, Biostars (<https://www.biostars.org/>) provides a very useful medium where questions and answers are exchanged amongst bioinformaticians.

Parameters used in BWA for read alignments:

```
bwa aln -o 1 -e 50 -m 10000 -t 4 -i 15 -q 10 -I (-I at the end is for Illumina NGS platforms)
```

Parameters used in GATK (for SNPs):

```
java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -stand_call_conf 50 -stand_emit_conf 10.0 -A DepthOfCoverage -A RMSMappingQuality -baq CALCULATE_AS_NECESSARY
```

Parameters used in GATK (for InDels):

```
java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -stand_call_conf 50 -stand_emit_conf 10.0 -A DepthOfCoverage -A RMSMappingQuality -baq CALCULATE_AS_NECESSARY -glm INDEL
```

Obtaining Ensembl VEP annotations for VCFs (including SIFT, Polyphen and Condel predictions):

- 1- Download latest package (and *plugins) from Ensembl website:
(www.ensembl.org/info/docs/variation/vep/index.html)
- 2- Tar xvf downloaded file(s)
- 3- perl INSTALL.pl – and download *Homo sapiens* cache(s)
- 4- perl variant_effect_predictor.pl -i **file.vcf** -o **file.vep** --protein --cache --regulatory --gmaf --force_overwrite --sift b --polyphen b --plugin Condel,/data/home/~/**ensembl-tools-release-75/scripts/variant_effect_predictor/ensembl-variation-VEP_plugins-e6cec6a/config**/Condel/config.b --fork 8 --canonical --individual all --pubmed --maf_esp --symbol

*to use Condel plugin:

1- Download latest Ensembl plugins from: https://github.com/ensembl-variation/VEP_plugins

2- tar -xvf downloaded file

2- mv Condel.pm ~/.vep/Plugins (create Plugins folder if not there; also .vep is a hidden folder)

3- edit the condel_SP.conf file (in config/Condel/config/) and set the 'condel.dir' parameter to `/data/home/~/.variant_effect_predictor/ensembl-variation-VEP_plugins-e6cec6a/config/Condel`

Example of commands used to filter variants in VEP file: To grab list of all rare/unique and homozygous mutations in candidate genes:

```
grep -f Candidate_genes.txt file.vep | grep -f PHI_SO_terms.txt | grep CANONICAL | grep HOM | grep _[A-Z]/ > file_candidate_mutations.txt
```

or use `grep GMAF=[A-Z]:0.00` instead of `grep _[A-Z]/` for variants which are present in the 1000GP but rare

Files used:

Candidate_genes.txt: a text file containing Ensembl IDs of your candidate genes – one per row

PHI_SO_terms.txt: a text file containing Ensembl VEP's SO terms which would be classified as a Φ mutation (available in section 9.8)

Command for Autozygosity plotting in AutoZplotter: `python autozplotter.py`

Parameters used for Autozygosity mapping in Plink: `plink --file <ped/map> --homozyg --noweb --homozyg-window-kb 1000 --homozyg-window-het 1 --homozyg-group --out <output>`

9.5. Appendices for Chapters 4 and 6 – Families with PCD

9.5.1. PCD sample quality

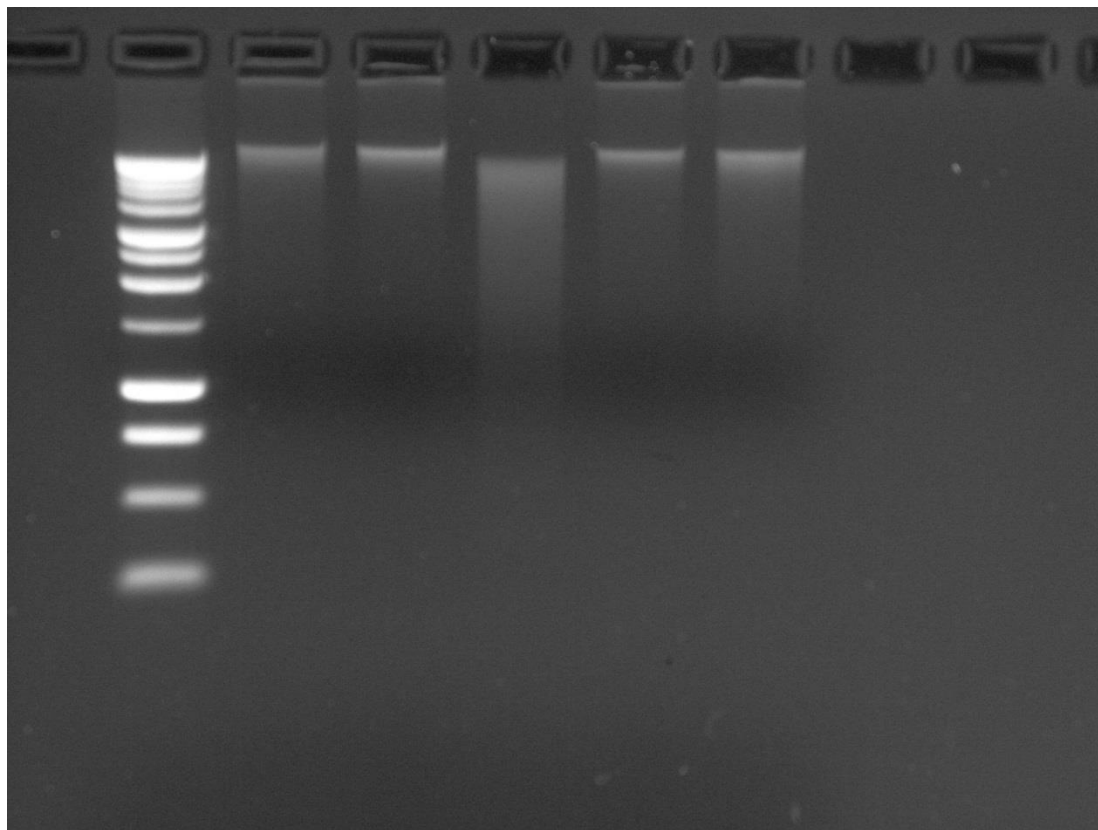


Figure 10.4 PCD Sample 5 tested for integrity on 19/09/13 before sending to BGI-Tech (even though highly degraded, some intact DNA could still be observed thus was sent in very large proportions to save the sample).

9.5.2. All known and potential PCD causal genes

The table below (i.e. Table 10.1) lists all the known human PCD causal genes and the variants which cause them. The table also includes genes which have been studied in model organisms such as Zebrafish, mice and *Chlamydomonas* which have resulted in ciliary dysfunction in these species implicating them as potential human PCD causal genes. The phenotypes caused as a result of the mutation, other information relating to the phenotype(s) and the references are also included below. The list was compiled with the help of Aasiya Ginwalla at the University of Bristol.

Gene, causal mutation	Species	Phenotype	Other information	References/Notes
RSPH4A , c.166dup, (p.Arg56Profs*11)	<i>Homo sapiens</i>	Transposition defect, radial-spoke head protein involved in ciliary movement	a radial-spoke head protein involved in ciliary movement	Unexpected genetic heterogeneity for primary ciliary dyskinesia in the Irish Traveller population.[507] Couldn't access paper, abstract info only.
CCNO , c.258_262dup, (p.Gln88Argfs*8) (other possible candidate genes <i>KCNN3</i> , <i>CDKN1C</i>)	<i>Homo sapiens</i>	Nude epithelium. Ciliary aplasia.	DNA repair gene involved in multiciliogenesis.	
DYX1C1 , 3.5 kb deletion	<i>Homo sapiens</i>	Absence of IDA and ODA	Neuronal migration gene	
RSPH1 , c.275-2A>C (homozygous splice variant), c.85G>C (p.Glu29*) and c.407_410delAGTC (p.Lys136Metfs*6)	<i>Homo sapiens</i>	Abnormal circular beat pattern but normal beat frequency, milder respiratory disease, higher NO levels than 'classic' PCD.	See "Knowles data supplement.pdf"	Mutations in RSPH1 cause primary ciliary dyskinesia with a unique clinical and ciliary phenotype.[408]
RSPH4A . No causal mutation mentioned	<i>Homo sapiens</i>	Radial spoke head proteins affected. Circular/elliptical motion. Normal beat frequency. Absence of central pair and transposed cilia may be seen.		Characterizing the ultrastructure of primary ciliary dyskinesia transposition defect using electron tomography.[508] IMAGES AVAILABLE
RSPH4A , c.325C>T (p.Q109X); c.1068G>A (p.V356X); c.IVS3+2-5(TAGG)del (mutation in conserved donor splice site); c.1468C>T (p.R490X); c.IVS5-4A>G; 3'UTR+195-205(11bp)del	<i>Homo sapiens</i>	Defects of microtubule (MT) organisation. Most frequent defect: absence of central pair. Ciliary disorientation, with a discordant alignment of the central pair in neighbouring cilia. Radial spokes, nexin links and IDA were indistinguishable, even in cross sections with normal MT pattern – could be due to mutation, or could reflect inadequate quality of specimens.		Mutations in radial spoke head genes and ultrastructural cilia defects in East-European cohort of primary ciliary dyskinesia patients.[509] IMAGES AVAILABLE
RSPH9 (aka c6orf206), c.801_803delGAA (p.Lys268)	<i>Homo sapiens</i>	Intermittent loss of central pair, seen occasionally in longitudinal section	Intermittent loss thought to be due to less stable central microtubules. Mutation likely to be hypomorphic.	Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia

del); c.804_806delGAA (p.Lys268 del)				with central-microtubular-pair abnormalities.[274] IMAGES AVAILABLE
RSPH4A , c.325C→T (p.Gln154X); c.1468C→T (p.Arg490X) c.460C→T (p.Gln154X); c.259C→T (p.Pro87Ser)	<i>Homo sapiens</i>	Disruption of the “radial spoke domain” likely. Truncation associated with loss of central pair		
RSP9 , c.780-783delCGC (p.Arg261del)	<i>Chlamydomonas</i>	Immotile	Mimics this human mutation c.801_803delGAA (p.Lys268del)	
<i>pf17</i> mutant, has mutation in RSP9 : c.131delG (p.Ser45AlafsX3)	<i>Chlamydomonas</i>	Immotile; radial spoke head complex is absent, and CP displacement, rather than loss.	Ortholog of human RSPH9	
c15orf26 knockdown	Zebrafish and planaria	ODA assembly blocked		Zebrafish Ciliopathy Screen Plus Human Mutational Analysis Identifies C21orf59 and CCDC65 Defects as Causing Primary Ciliary Dyskinesia.[407] IMAGES AVAILABLE
ccdc65 knockdown	Zebrafish and planaria	Cilia beat pattern altered	<i>Chlamydomonas</i> ida6 mutant identifies CCDC65/FAP250 as an essential component of the nexin-dynein regulatory complex	
Ccdc65 , c.877_878delAT (p.Ile293Profs*2)	<i>Homo sapiens</i>	Normal ODA, radial spokes, and central pairs but a reduction in IDA and nexin links. Some microtubule disorganization. Stiff, dyskinetic cilia. Recurrent bronchitis, sinusitis, and/or otitis media		
C21orf59 knockdown	Zebrafish and planaria	ODA assembly blocked; loss of both DA components, immotile cilia.	C21orf59 <i>Chlamydomonas</i> ortholog, FBB18, is a flagellar matrix protein that accumulates specifically when cilia motility is impaired	
C21orf59 , truncating mutations: c.292C>T (p.Arg98*), c.735C>G (p.Tyr245*), and c.792_795delTTTA (p.Tyr264*)	<i>Homo sapiens</i>	Truncating mutations lead to: loss of both DA, immotile cilia in some cases		

Missense variants: c.97C>T (p.Arg33Trp), c.422A>G (p.Asp141Gly), and c.517G>T (p.Asp173Thr)		Missense variants lead to: low nasal nitric oxide, sinus disease, otitis media, and/or bronchiectasis but normal ciliary ultrastructure.	Suggests that these mutations might be either hypomorphic or not causative of PCD.	
RSPH1 , c.281G>A (p.Trp94*); c.275-2A>C splice site change c.85G>T (p.Glu29*)	<i>Homo sapiens</i>	Transposition of peripheral outer microtubules into the 'empty' CP space, accompanied by a distinctive intermittent loss of the central pair microtubules		Targeted NGS gene panel identifies mutations in RSPH1 causing primary ciliary dyskinesia and a common mechanism for ciliary central pair agenesis due to radial spoke defects.[510] IMAGES AVAILABLE
DNAH12 , c.5093G>A (p.Pro1698Leu) c.3577C>A (p.Ala1193Ser) (rs1511075)	<i>Homo sapiens</i>			
DNAH1 , c.3103C>T (p.Arg1035Cys) and c.12090 + 7A>C splice site	<i>Homo sapiens</i>			
DNAH3 , c.5368T>A (p.Ile1790Phe) and c.7C>T (p.Ala3Thr)	<i>Homo sapiens</i>			
WDR66 , c.186-187ins of 15 base-pairs (p.62-63insEEEEK)	<i>Homo sapiens</i>			
CCDC40 , c.3040-3041insCAC (p.1014insThr)	<i>Homo sapiens</i>			
DNAH5 , c.5563insA (p.1855NfsX5); c.8440delGAACCAAA (p.2814fsX1); c.4361G→A and c.8910+8911delAT→insG (p.R1454Q and p.2970SfsX7);	<i>Homo sapiens</i>	Absence of ODA is observed on all peripheral doublets. Randomization of left-right asymmetry.		Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry.[371] IMAGES AVAILABLE

c.10555G→C (p.G3519R); c.1VS74-1G→C (splice-site mutation); c.4360C→T and [?] (p.R1454X and [?]); c.7915C→T and [?] (p.R2639X and [?]); c.1828C→T and c.5130insA (p.Q610X and p.R1711TfsX36)				
DNAH5 , See Table 1 in paper	<i>Homo sapiens</i>	Absence of ODA. Correct assembly of ODA complexes affected.		DNAH5 Mutations Are a Common Cause of Primary Ciliary Dyskinesia with Outer Dynein Arm Defects.[372]
DNAH5 , c.5563insA, c.8440delAACC AAA. c.IVS74-1G > C	<i>Homo sapiens</i>	Immotile, ODA defect Slow beat frequency and uncoordinated		Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia.[511]
DNAH5 , c.7829A>G (p.Glu2610Gly)	<i>Homo sapiens</i>	Asplenia syndrome, often seen in PCD patients.		A novel mutation of the axonemal dynein heavy chain gene 5 (DNAH5) in a Japanese neonate with asplenia syndrome.[512]
DNAI2 , loss of function, IVS11+1G > A	<i>Homo sapiens</i>	ODA defects; lack of DNAI2 protein expression	Skipping of exon 11 (IVS11+1G > A)	DNAI2 Mutations Cause Primary Ciliary Dyskinesia with Defects in the Outer Dynein Arm.[382] IMAGES AVAILABLE
DNAI2 , nonsense mutation (c.787C > T)	<i>Homo sapiens</i>	ODA defects; truncated protein	(IVS3-3T > G) Splicing mutation causes skipping of exon 4 and a frame shift that results in a premature stop codon (I116GfsX54)	
DNAI2 , splicing mutation (IVS3- 3T > G)	<i>Homo sapiens</i>	ODA defects	Absence of the ODA heavy chains DNAH5 and DNAH9 from all DNAI2 mutant ciliary axonemes. DNAI2 mutations affect assembly	

			of proximal and distal ODA complexes	
<i>DNAL1</i> , homozygous point mutation c.449A>G; p.Asn150Ser	<i>Homo sapiens</i>	Absent or markedly shortened ODA. Reduced stability of the axonemal dynein light chain 1 and damaged interactions with dynein heavy chain and with tubulin.	Asn at position150, is critical for the proper tight turn between the β strand and the α helix of the leucine-rich repeat in the hydrophobic face that connects to the dynein heavy chain.	Primary Ciliary Dyskinesia caused by homozygous mutation in DNAL1, encoding dynein light chain 1.[398] IMAGES AVAILABLE
<i>DNAL1</i> ,				Identification and Analysis of Axonemal Dynein Light Chain 1 in Primary Ciliary Dyskinesia Patients.
<i>LRRC50</i> , p.Leu175Arg; c.1349_1350insC) with (p.Pro451AlafsX5); p.Arg271X;	<i>Homo sapiens</i>	Disrupts SDS22-like subfamily LRR structure	This LRR might participate in the structural link between the inner- and outer-row dyneins during their preassembly	Deletions and Point Mutations of <i>LRRC50</i> Cause Primary Ciliary Dyskinesia Due to Dynein Arm Defects.[396]
<i>CCDC151</i> . Mutation not yet found in humans.	Zebrafish	Deregulated ciliary length	Assembly of motile cilia	The coiled-coil domain containing protein CCDC151 is required for the function of IFT-dependent motile cilia in animals.[513]
<i>Spag17</i> . Mutation not yet found in humans	Mice	Immotile cilia; reduced bridge density and greater separation between the two CP microtubules and the absence of a C1 projection		Sperm-associated antigen-17 gene is essential for motile cilia function and neonatal survival.[514] IMAGES AVAILABLE
USEFUL ARTICLE				Ciliary Genes Are Down-Regulated in Bronchial Tissue of Primary Ciliary Dyskinesia Patients. [515]
<i>DNAH11</i> , R2852X. Mouse model <i>iv/iv</i> , gene <i>dnah11</i>	<i>Homo sapiens</i> ; mice	<i>Situs inversus totalis</i> ; immotile cilia in embryonic node.	Also an R3004Q substitution in a conserved position	Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of <i>situs</i>

				<i>inversus totalis</i> and most likely primary ciliary dyskinesia.[516]
DNAH11 , c.12384C>G (p.Y4128X) and c.13552_13608del (p.A4518_A4523delinsQ) c.11059A>G; p.K3687E	<i>Homo sapiens</i>	Normal ultrastructure; reduced waveform amplitude and hyperkinetic beating pattern, so defective function.		Primary ciliary dyskinesia associated with normal axoneme ultrastructure is caused by DNAH11 mutations.[397] IMAGES AVAILABLE
	<i>Homo sapiens; Chlamydomonas</i>	β -DHC can assemble outer arm subunits into the flagellar axoneme, but swimming velocity and/or beat frequency are reduced		Mutations of DNAH11 in Primary Ciliary Dyskinesia Patients with Normal Ciliary Ultrastructure.[381]
	Mouse, <i>dnahc11</i> , iv/iv model	<i>situs</i> defects; no detectable ciliary beat frequency, and suffer otitis media and rhinitis, even though they have normal ciliary ultrastructure		
DNAI1	<i>Homo sapiens</i>	Absence of ODA		Loss-of-Function Mutations in a Human Gene Related to <i>Chlamydomonas reinhardtii</i> Dynein IC78 Result in Primary Ciliary Dyskinesia.[376] IMAGES AVAILABLE
DNAI1 , c.219+3insT/W568X	<i>Homo sapiens</i>	Slow beat frequency, uncoordinated movement		Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia.[511]
TXNDC3 , nonsense Mutation (p.Leu426X) and a Common Intronic Variant (c.271–27C>T)	<i>Homo sapiens</i>	ODA defects	Thioredoxin family	A common variant in combination with a nonsense mutation in a member of the thioredoxin family

				causes primary ciliary dyskinesia.[385] IMAGES AVAILABLE
ktu , c.841C>A (p.Y243X)	Medaka (<i>Oryzias latipes</i>)	Partial or complete loss of ODA and IDA.		Ktu/PF13 is required for cytoplasmic pre-assembly of axonemal dyneins.[384] IMAGES AVAILABLE
KTU (aka DNAAF2), c.1214^1215insACGATACCTG CGTGGC (p.G406Rfs89X); c.23C>A (p.S8X)	<i>Homo sapiens</i>	Assembly of distally located type 2 ODA complexes affected, which specifically contain dynein β -HC (heavy chain) orthologue DNAH9	PF13 is KTU homologue in <i>Chlamydomonas</i>	
LRRC50 (aka DNAAF1), c.792C>A, (p.Tyr264X) nonsense mutation	<i>Homo sapiens</i>	Severely truncated protein, which is involved in ODA assembly and stability causing IDA defects too.		Loss-of-function mutations in the human ortholog of <i>Chlamydomonas reinhardtii</i> ODA7 disrupt dynein arm assembly and cause primary ciliary dyskinesia.[377] IMAGES AVAILABLE
SPAG1 c.2T>G (p.Met1?)	<i>Homo sapiens</i>	Defective ODA and IDA		Mutations in SPAG1 cause primary ciliary dyskinesia associated with defective outer and inner dynein arms.[406]
CCDC39 ; c.1168-32A>G (p.Arg96X) p.Glu731AsnfsX31 p.Thr358GlnfsX3, p.Ser786IlefsX33	<i>Homo sapiens</i> and Old English Sheepdog (Bobtail)	Assembly of IDA	FAP59 (<i>Chlamydomonas</i> orthologue)	CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs.[378] IMAGES AVAILABLE
CCDC40 ; c.C1366T (p.R449X)	<i>Homo sapiens</i>	Severe cilia beating defects	Table with CCD40 mutations available	The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation.[380] IMAGES AVAILABLE
CCDC103 ; c.383_384insG (p.Gly128fs25*) p.His154Pro	<i>Homo sapiens</i>	Reduced ODA, assembly and stability affected, causing accompanying IDA defects.	Ccdc103 is a dynein arm attachment factor	CCDC103 mutations cause primary ciliary dyskinesia by disrupting

				assembly of ciliary dynein arms.[386] IMAGES AVAILABLE
HEATR2 ; p.Leu795Pro	<i>Homo sapiens</i>	Reduced HEATR2 levels, absent dynein arms, and loss of ciliary beating.	Suggests role in either dynein arm transport or assembly.	Whole-exome capture and sequencing identifies HEATR2 mutation as a cause of primary ciliary dyskinesia.[399] IMAGES AVAILABLE
	<i>Chlamydomonas reinhardtii</i>	Absent ODA, reduced flagellar beat frequency, and decreased cell velocity		
LRRC6 , c.220G>C (p.Ala24Pro); c.436G>C (p.Asp146His); c.574C>T (p.Gln192*); c.576dupA (p.Glu193Argfs*4); c.598-599delAA (p.Lys200Glufs*3)	<i>Homo sapiens</i>	Absence of Both DAs in Respiratory Cilia and/or Spermatozoa Flagella Possible <i>situs inversus</i>	Similar to DNAAF1	Loss-of-function mutations in LRRC6, a gene essential for proper axonemal assembly of inner and outer dynein arms, cause primary ciliary dyskinesia.[383] IMAGES AVAILABLE
DNAAF3 (PF22) ; c.323T>C (p.Leu108Pro); c.406C>T (p.Arg136X); c.762_763insT (p.Val255CysfsX12); c.973G>A (p.Ala325Thr)	<i>Homo sapiens</i>	<i>Situs inversus</i> ; defects in assembly of IDA and ODA.	DNAAF3 is important for the assembly of outer and inner dynein arms along the entire length of the axoneme, comprising ODA types 1 and 2 and the DNALI1 containing IDA types.	Mutations in axonemal dynein assembly factor DNAAF3 cause primary ciliary dyskinesia.[34] IMAGES AVAILABLE
	Zebrafish	Zebrafish <i>dnaaf3</i> knockdown also disrupts dynein arm assembly and ciliary motility, causing PCD phenotypes e.g. hydrocephalus and laterality malformations.		
	<i>Chlamydomonas reinhardtii</i>	<i>Chlamydomonas reinhardtii</i> PF22 is exclusively cytoplasmic, and a null mutant fails to assemble ODA and some IDA.		
HYDIN : c.3985G>T; c.922A>T (p.Lys307*)	<i>Homo sapiens</i>	Normal 9+2 axonemal composition, DAs unaffected. 9+0 cilia and 8+1 cilia (ciliary	Mostly normal ultrastructure.	Recessive HYDIN mutations cause primary ciliary dyskinesia without

		transposition defect where CP is absent and replaced by one of the peripheral microtubules) were only rarely observed. A few 9+3 cilia observed. Abnormal axonemal bending.		randomization of left-right body asymmetry.[517] IMAGES AVAILABLE
CCDC114 ; c.742G>A	<i>Homo sapiens</i> <i>Chlamydomonas</i> DCC2	Complete absence of ciliary ODAs, so immotile cilia.	CCDC114 is an essential ciliary protein required for microtubular attachment of ODAs Fertility seems not to be greatly affected by CCDC114 deficiency	Splice-Site Mutations in the Axonemal Outer Dynein Arm Docking Complex Gene CCDC114 Cause Primary Ciliary Dyskinesia.[518]
CCDC164 , c.2056A>T (p.Lys686*); p.Gln118*	<i>Homo sapiens</i>	Defects in assembly of the N-DRC structure and closely associated IDA, and defective ciliary movement	Gene encodes DRC1. No <i>situs inversus</i> in humans, but suggestions of this in mice.	The nexin-dynein regulatory complex subunit DRC1 is essential for motile cilia function in algae and humans.[389] IMAGES AVAILABLE
	<i>Chlamydomonas</i>	Reduced swimming speed and abnormal ciliary waveform		
CCDC65	<i>Homo sapiens</i> ; <i>Chlamydomonas</i>	Stiff and dyskinetic cilia beating patterns, no detectable ultrastructural defects of the ciliary axoneme	A nexin-dynein regulatory complex member Gas8, a nexin-dynein regulatory complex component previously identified to associate with CCDC65, was absent	CCDC65 mutation causes primary ciliary dyskinesia with normal ultrastructure and hyperkinetic cilia.[519] IMAGES AVAILABLE
ZMYND10 (aka BLU), c.47T>G (p.Val16Gly); c.589_590del; c.797T>C (p.Leu266Pro); c.116T>C (p.Leu39Pro); c.65delT (p.Phe22Serfs*21)	<i>Homo sapiens</i>	Absence of DA, immotile cilia. C.47T>G special case: stiff and slowed beat, as DA partially retained	ZMYND10 is a cytoplasmic protein required for IDA and ODA assembly. Its variants cause ciliary dysmotility and PCD with laterality defects.	Mutations in ZMYND10, a gene essential for proper axonemal assembly of inner and outer dynein arms in humans and flies, cause primary ciliary dyskinesia.[520] IMAGES AVAILABLE
	<i>Drosophila</i>	IDA and ODA defects, proprioception deficits, and sterility due to immotile sperm		

ZMYND10 , c.1136A>G (p.Tyr379Cys), c.630delG (seems to be founder mutation), c.169_173delinsTCCCAAT (p.Gly57Serfs*3), c.[259T>C];[436G>C] (p.[Cys87Arg];[Asp146His]), c.562C>T (p.Gln188*), c.630delG (p.Trp210Cysfs*12), c.598_599delAA (p.Lys200Glu fs*3), c.653+1G>A, c.710_711delCA (p.Thr237Lysfs*7), and c.891delA (p.Ala298Profs*2)	<i>Homo sapiens</i>	ODA and IDA defects		ZMYND10 is mutated in primary ciliary dyskinesia and interacts with LRRC6.[405]
	Zebrafish	Ciliary paralysis, leading to cystic kidneys and otolith defects		
	Xenopus	Interference with ciliogenesis		
RSPH1 , c.85G>T (p.Glu29*); c.407_410delAGTA (p.Lys136Metfs*6); c.308G>A (p.Gly103Asp); splice site mutations: c.275-2A>C, c.366-3C>A, c.727+5G>A. Also c.366G>A [p.=]	<i>Homo sapiens</i>	Sinopulmonary syndrome and <i>situs solitus</i> , fertility problems. Cilia with different beating patterns (active, slow, or immotile; abnormal beating patterns and reduced ciliary beat frequencies.	Absence of RSPH1 leads to an abnormal axonemal configuration with CC defects and an absence of RSs in cilia with no CCs	Loss-of-function mutations in RSPH1 cause primary ciliary dyskinesia with central-complex and radial-spoke defects.[35] IMAGES AVAILABLE
ARMC4 , c.2675C>A (p.Ser892*) and c.1972G>T (p.Glu658*)	<i>Homo sapiens</i>	ARMC4 deficiency, loss of the distal ODA motors responsible for generating ciliary beating, giving rise to cilia immotility. <i>Situs inversus</i>	ARMC4 gene expression is upregulated during ciliogenesis. Predicts interaction with the DNAI2.	Combined exome and whole-genome sequencing identifies mutations in ARMC4 as a cause of primary ciliary dyskinesia with defects in the outer dynein arm.[521] IMAGES AVAILABLE
DYX1C1 (aka DNAAF4)	Mice (c.T2A)	Severe hydrocephalus; <i>situs inversus totalis</i> , <i>situs ambiguous</i> or <i>situs solitus</i> .	A dynein axonemal assembly factor. Interacts with the	DYX1C1 is required for axonemal dynein assembly and ciliary motility.[522]

	Mouse ependymal cells	<i>Lacking ODA heavy chain Mdnah5, and IDA light chain, Dnali1</i>	cytoplasmic ODA/IDA assembly factor DNAAF2/KTU	IMAGES AVAILABLE
	Mouse tracheal cells	<i>Absent DA, normal 9+2 and intact radial spokes</i>		
	Mouse respiratory cells	<i>Absent DA subunits</i>		
	Zebrafish	Body curvature, hydrocephalus, cystic kidneys and <i>situs inversus</i>		
DYX1C1 , c.253_254insGA (p.T85Rfs4*); c.325G>T (p.E109*); c.384C>A (p.Y128*); c.390_393delAAGT (p.V132*); c.485G>A (p.W162*); c.583delA (p.I195*); c.783+1G>T; c.808C>T (p.R270*)	<i>Homo sapiens</i>	DA defects and immotile cilia. Absence of proteins found in both DA subunits. <i>Situs inversus</i> , <i>ambiguous</i> or <i>solitus</i> .	DYX1C1 is similar to KTU/DNAAF2 and interacts with it at an early step of cytoplasmic ODA and IDA assembly.	
DNAH1 , c.11788-1G>A (p.Gly3930Alafs*120); c.3877G>A (p.Asp1293Asn); c.12796 T>C (p.4266Glnext*21); c.5094+1G>A (p.Leu1700Serfs72);	<i>Homo sapiens</i>	Impaired sperm motility due to absent, short, coiled, bent, and irregular flagella; reduced sperm motility. IDA mostly absent; one third of the microtubule doublets were malformed or absent; central singlet of microtubules was missing (9+0) in half of cases; fibrous sheath was strongly disorganized.		Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella.[523] IMAGES AVAILABLE
RPGR	<i>Homo sapiens</i>	<i>Retinitis pigmentosa</i> ; complex dynein arm defect		Ciliary defects and genetics of primary ciliary dyskinesia.[524]
DPCD ; possibly 168T>G responsible	Mouse	Defect in IDA or defect in both DA	Data suggest that DPCD plays a role in the formation or function of ciliated cells	Investigation of the possible role of a novel gene, DPCD, in primary ciliary dyskinesia.[525]

Table 10.1 All known human PCD causal genes and the variants which cause them. Where images and/or videos from cilia were available, it was noted.

9.5.3. Looking for causal variants

- 1- To obtain a list of all the mutations in known candidate genes:

```
grep -f Human_PCD_genes.txt file.vep | grep -f PHI_SO_terms.txt | grep CANONICAL |  
grep -v synonymous_variant | grep -v intron_variant | grep _[A-Z]/ >  
file_known_PCD_gene_mutations.txt
```

or use 'GMAF=[A-Z]:0.00' instead of 'grep _[A-Z]/' if initial analysis is fruitless

- 2- To obtain a list of all the mutations in other candidate genes:

```
grep -f Suspected_Ciliome_genes.txt 1.vep | grep -f PHI_SO_terms.txt | grep  
CANONICAL | grep -v synonymous_variant | grep -v intron_variant | grep _[A-Z]/ >  
file_suspected_PCD_gene_mutations.txt
```

or use 'GMAF=[A-Z]:0.00' instead of 'grep _[A-Z]/' if initial analysis is fruitless

- 3- To obtain a list of all the candidate genes' location*:

python PCD_candid_genes_location.py (change name of files for 'known' and 'suspected')

- 4- To find out which of these mutations are homozygous (for recessive mutations):

```
grep -f file_suspected_PCD_gene_location.txt file.vcf | grep 1/1 >  
file_suspected_PCD_gene_homozygote_mutations.txt
```

- 5- ***/****Convert resulting files to Stata format (e.g. import text data) and merge

```
file_suspected_PCD_gene_homozygote_mutations_header.dta and  
file_suspected_PCD_gene_mutations_header.dta together (using 'rsid' and 'minor'  
as key variables) > file_merged_homoz_susp_candid_mutations.dta
```

- 6- Analyze file_merged_homoz_candid_mutations.dta files

- Check for notable stop gains and frameshifting indels
- Check whether mutation is in a autozygous region
- Delete (from proband) homozygote mutations present in control(s)
- Check quality of base calls
- Remove ones with GMAF > 0.01 in 1000GP
- Use STRING for interactome
- Check GERP score
- Check UK10K & EVS for presence and MAF

- 7- Repeating the above for frameshift indels which are homozygous using:

* The python script can be found in section 10.5.4


```
grep -f SO_terms_InDel.txt file.indel.annot.csv | grep 1/1 | grep -v rs | cat >
file_cand_homoz_indel.txt
```

Steps 4 and 5 have now become obsolete with the ‘--individual all’ option provided by Ensembl VEP v80. This option also adds the ‘HOM’ (or HET) flag for individuals who are homozygotes (or heterozygotes) for the variant allele. Step 7 is also no longer needed with the inclusion of indel related SO terms in the ‘PHI_SO_terms.txt’ file.

9.5.4. Files used in PCD analyses

These files contain already identified and suspected PCD genes (their Ensembl IDs) respectively. The first file was generated manually using the literature; and the second one was generated with the help of scripts using the Ciliome database (available at: http://www.sfu.ca/~leroux/ciliome_home.htm).

Human_PCD_genes.txt

```
ENSG00000039139
ENSG00000122735
ENSG00000167646
ENSG00000154099
ENSG00000086288
ENSG00000105877
ENSG00000171595
ENSG00000165506
ENSG00000111834
ENSG00000172426
ENSG00000145075
ENSG00000141519
ENSG00000119661
ENSG00000167131
ENSG00000129295
ENSG00000157423
ENSG00000164818
ENSG00000046651
ENSG00000156313
ENSG00000105479
ENSG00000157856
ENSG00000139537
ENSG00000160188
ENSG00000256061
ENSG00000004838
```

Suspected_Ciliome_genes.txt

ENSG00000256061	ENSG00000163913	ENSG00000183690	ENSG00000197958	ENSG00000197748	ENSG00000112992	ENSG00000141342
ENSG00000004838	ENSG00000138160	ENSG00000100744	ENSG00000187778	ENSG00000165990	ENSG00000142168	ENSG00000122435
ENSG00000101052	ENSG00000090863	ENSG00000169299	ENSG00000197894	ENSG00000104983	ENSG00000176619	ENSG00000185246
ENSG00000114446	ENSG00000065183	ENSG00000085415	ENSG00000107223	ENSG00000165695	ENSG00000147400	ENSG00000104047
ENSG00000163879	ENSG00000160226	ENSG00000130348	ENSG00000123454	ENSG00000132321	ENSG00000169660	ENSG00000132437
ENSG00000068885	ENSG00000119927	ENSG00000077147	ENSG00000163378	ENSG00000099889	ENSG00000175110	ENSG00000140057
ENSG00000032742	ENSG00000152582	ENSG00000167972	ENSG00000100003	ENSG00000135318	ENSG00000108551	ENSG00000137522
ENSG00000138002	ENSG00000159079	ENSG00000010292	ENSG00000108592	ENSG00000088727	ENSG00000088247	ENSG00000186889
ENSG00000112530	ENSG00000141577	ENSG00000147576	ENSG00000115942	ENSG00000156787	ENSG00000095459	ENSG00000185008
ENSG00000145107	ENSG00000135205	ENSG00000168671	ENSG00000172409	ENSG00000103599	ENSG00000137161	ENSG00000134438
ENSG00000103351	ENSG00000151023	ENSG00000109133	ENSG00000089048	ENSG00000167815	ENSG00000145214	ENSG00000137413
ENSG00000163093	ENSG00000168625	ENSG00000198399	ENSG00000186298	ENSG00000087302	ENSG00000163714	ENSG00000095383
ENSG00000122970	ENSG00000166596	ENSG00000163001	ENSG00000038382	ENSG00000116198	ENSG00000150753	ENSG00000141570
ENSG00000122735	ENSG00000145491	ENSG00000134186	ENSG00000131242	ENSG00000106479	ENSG00000146872	ENSG00000151806
ENSG00000111834	ENSG00000164675	ENSG00000197930	ENSG00000159200	ENSG00000164012	ENSG00000171962	ENSG00000035059
ENSG00000171595	ENSG00000183833	ENSG00000148737	ENSG00000129347	ENSG00000156042	ENSG00000181085	ENSG00000113649
ENSG00000130363	ENSG00000163885	ENSG00000110756	ENSG00000164144	ENSG00000112210	ENSG00000122642	ENSG00000057468
ENSG00000141013	ENSG00000112981	ENSG00000142186	ENSG00000185009	ENSG00000119333	ENSG00000139318	ENSG00000187609
ENSG00000105948	ENSG00000161973	ENSG00000001036	ENSG00000064419	ENSG00000101882	ENSG00000173113	ENSG00000054116
ENSG00000138587	ENSG00000101222	ENSG00000146143	ENSG00000147604	ENSG00000166855	ENSG00000159556	ENSG00000111361
ENSG00000109083	ENSG00000103021	ENSG00000124193	ENSG00000105254	ENSG00000109618	ENSG00000168291	ENSG00000154035
ENSG00000128581	ENSG00000184154	ENSG00000166391	ENSG00000136485	ENSG00000143493	ENSG00000072786	ENSG00000103067
ENSG00000167858	ENSG00000163610	ENSG00000143995	ENSG00000090273	ENSG00000089177	ENSG00000137876	ENSG00000179292
ENSG00000111837	ENSG00000114841	ENSG00000135587	ENSG00000138678	ENSG00000181378	ENSG00000111880	ENSG00000135968
ENSG00000184009	ENSG00000080824	ENSG00000164885	ENSG00000165219	ENSG00000168385	ENSG00000048392	ENSG00000109536
ENSG00000108641	ENSG00000122507	ENSG00000188906	ENSG00000105568	ENSG00000119650	ENSG00000117713	ENSG00000155906
ENSG00000196659	ENSG00000173540	ENSG00000169359	ENSG00000196586	ENSG00000119640	ENSG00000154889	ENSG00000006717
ENSG00000118096	ENSG00000089101	ENSG00000125971	ENSG00000119698	ENSG00000166024	ENSG00000109339	ENSG00000113300
ENSG00000146425	ENSG00000158023	ENSG00000131873	ENSG00000110917	ENSG00000089248	ENSG00000115866	ENSG00000182858
ENSG00000172426	ENSG00000139428	ENSG00000147457	ENSG00000135636	ENSG00000115459	ENSG00000135414	ENSG00000154310
ENSG00000118997	ENSG00000007174	ENSG00000176749	ENSG00000090061	ENSG00000149084	ENSG00000095321	ENSG00000166206
ENSG00000039139	ENSG00000124074	ENSG00000012963	ENSG00000135472	ENSG00000110711	ENSG00000168028	ENSG00000164051
ENSG00000128408	ENSG00000096872	ENSG00000118873	ENSG00000162368	ENSG00000100031	ENSG00000145362	ENSG00000171863
ENSG00000096093	ENSG00000100360	ENSG00000143933	ENSG00000170889	ENSG00000162643	ENSG00000145782	ENSG00000175536
ENSG00000162961	ENSG00000146221	ENSG00000150457	ENSG00000125247	ENSG00000115953	ENSG00000121879	ENSG00000125877
ENSG00000157796	ENSG00000170385	ENSG00000117868	ENSG00000164587	ENSG00000167977	ENSG00000174444	ENSG00000160200
ENSG00000123810	ENSG00000176101	ENSG00000150316	ENSG00000175155	ENSG00000183828	ENSG00000075856	ENSG00000164329
ENSG00000165533	ENSG00000061918	ENSG00000179115	ENSG00000134313	ENSG00000157106	ENSG00000198718	ENSG00000132130
ENSG00000168589	ENSG00000099219	ENSG00000107937	ENSG00000189067	ENSG00000152977	ENSG00000128039	ENSG00000164815
ENSG00000084207	ENSG00000198730	ENSG00000131951	ENSG00000112514	ENSG00000111727	ENSG00000130508	ENSG00000091656
ENSG00000102218	ENSG00000165659	ENSG00000082068	ENSG00000105137	ENSG00000158234	ENSG00000131018	ENSG00000142798
ENSG00000109971	ENSG00000080608	ENSG00000171735	ENSG00000165152	ENSG00000179632	ENSG00000068784	ENSG00000182853
ENSG00000125124	ENSG00000129348	ENSG00000171316	ENSG00000146476	ENSG00000113448	ENSG00000128524	ENSG00000100997
ENSG00000138175	ENSG00000102125	ENSG00000075624	ENSG00000119865	ENSG00000178662	ENSG00000100567	ENSG00000073417
ENSG00000075945	ENSG00000152520	ENSG00000114480	ENSG00000171097	ENSG00000165280	ENSG00000136848	ENSG00000158062
ENSG00000138036	ENSG00000108094	ENSG00000198783	ENSG00000090861	ENSG00000154099	ENSG00000098195	ENSG00000099246
ENSG00000150628	ENSG00000163637	ENSG00000139116	ENSG00000172554	ENSG00000114473	ENSG00000117174	ENSG00000105220
ENSG00000066185	ENSG00000119929	ENSG00000124181	ENSG00000167658	ENSG00000179636	ENSG00000100413	ENSG00000101844
ENSG00000152763	ENSG00000188723	ENSG00000149187	ENSG00000156194	ENSG00000100246	ENSG00000148672	ENSG00000021776
ENSG00000173093	ENSG00000162441	ENSG00000164089	ENSG00000066136	ENSG00000164983	ENSG00000112759	ENSG00000071894
ENSG00000182224	ENSG00000151914	ENSG00000183597	ENSG00000154803	ENSG00000109323	ENSG00000151151	ENSG00000183291
ENSG00000119661	ENSG00000181610	ENSG00000101935	ENSG00000121068	ENSG00000135373	ENSG00000133703	ENSG00000178234
ENSG00000077327	ENSG00000106852	ENSG00000147224	ENSG00000155100	ENSG00000112541	ENSG00000104219	ENSG00000124383
ENSG00000173013	ENSG00000166165	ENSG00000118689	ENSG00000086827	ENSG00000081870	ENSG00000163945	ENSG00000102910
ENSG00000143156	ENSG00000011566	ENSG00000165097	ENSG00000151414	ENSG00000159720	ENSG00000082898	ENSG00000196262
ENSG00000158710	ENSG00000043514	ENSG00000117475	ENSG00000186676	ENSG00000146282	ENSG00000149577	ENSG00000080546
ENSG00000118965	ENSG00000152683	ENSG00000147972	ENSG00000100220	ENSG00000104321	ENSG00000141429	ENSG00000084623
ENSG00000070761	ENSG00000104907	ENSG00000132466	ENSG00000135018	ENSG00000100401	ENSG00000180185	ENSG00000116299
ENSG00000167552	ENSG00000054392	ENSG00000068394	ENSG00000008300	ENSG00000126226	ENSG00000104723	ENSG00000107593
ENSG00000113966	ENSG00000069974	ENSG00000168439	ENSG00000129159	ENSG00000120738	ENSG00000137992	ENSG00000135241
ENSG00000159713	ENSG00000013455	ENSG00000093217	ENSG00000127152	ENSG00000166183	ENSG00000137962	ENSG00000183780
ENSG00000187535	ENSG00000137349	ENSG00000122484	ENSG00000171435	ENSG00000142609	ENSG00000129187	ENSG00000130713
ENSG00000131233	ENSG00000130283	ENSG00000107036	ENSG00000182197	ENSG00000121350	ENSG00000127022	ENSG00000184408
ENSG00000185055	ENSG00000086758	ENSG00000079739	ENSG00000116833	ENSG00000115423	ENSG00000178802	ENSG00000183576
ENSG00000188229	ENSG00000196531	ENSG00000102898	ENSG00000117593	ENSG00000066382	ENSG00000163798	ENSG00000178952
ENSG00000123607	ENSG00000134265	ENSG00000167986	ENSG00000166402	ENSG00000132514	ENSG00000076003	ENSG00000165782
ENSG00000138686	ENSG00000198416	ENSG00000163655	ENSG00000060688	ENSG00000185760	ENSG00000117139	ENSG00000073584
ENSG00000173597	ENSG00000164252	ENSG00000198900	ENSG00000172046	ENSG00000141378	ENSG00000172009	ENSG00000124228
ENSG00000121083	ENSG00000163312	ENSG00000116030	ENSG00000085978	ENSG00000090054	ENSG00000164073	ENSG00000129625
ENSG00000158104	ENSG00000177889	ENSG00000168348	ENSG00000105877	ENSG00000149273	ENSG00000163788	ENSG00000144559
ENSG00000158079	ENSG00000180957	ENSG00000183048	ENSG00000105519	ENSG00000196118	ENSG00000198722	ENSG00000126698
ENSG00000154678	ENSG00000160967	ENSG00000145349	ENSG00000108953	ENSG00000035115	ENSG00000104888	ENSG00000198755
ENSG00000116885	ENSG00000091157	ENSG00000198626	ENSG00000182768	ENSG00000008952	ENSG00000148396	ENSG00000066651
ENSG00000109103	ENSG00000125875	ENSG00000182749	ENSG00000008382	ENSG00000173349	ENSG00000111364	ENSG00000140455
ENSG00000048432	ENSG00000126934	ENSG00000108176	ENSG00000122507	ENSG00000117395	ENSG00000071051	ENSG00000104331
ENSG00000164953	ENSG00000102221	ENSG00000163516	ENSG00000160401	ENSG00000119636	ENSG00000166311	ENSG00000166971
ENSG00000132005	ENSG00000187919	ENSG00000138081	ENSG00000197826	ENSG00000100129	ENSG00000182687	ENSG00000123395
ENSG00000150995	ENSG00000137200	ENSG00000136944	ENSG00000010626	ENSG00000079335	ENSG00000164934	ENSG00000053328
ENSG00000164109	ENSG00000197045	ENSG00000143416	ENSG00000177076	ENSG00000157212	ENSG00000133313	
ENSG00000177963	ENSG00000154380	ENSG00000188878	ENSG00000181704	ENSG00000168906	ENSG00000163106	
ENSG00000160799	ENSG00000163113	ENSG00000171004	ENSG00000068323	ENSG00000169255	ENSG00000164062	
ENSG00000165629	ENSG00000113141	ENSG00000100296	ENSG00000173020	ENSG00000008740	ENSG00000169738	
ENSG00000027001	ENSG00000165813	ENSG00000090060	ENSG00000175220	ENSG00000177479	ENSG00000096717	

ENSG00000168090	ENSG00000011426	ENSG00000165671	ENSG00000005075	ENSG00000100380	ENSG00000100889	ENSG00000115306
ENSG00000128607	ENSG00000114573	ENSG00000130731	ENSG00000111186	ENSG00000168530	ENSG00000166341	ENSG00000163873
ENSG00000188419	ENSG00000134255	ENSG00000178105	ENSG00000055044	ENSG00000177093	ENSG00000153944	ENSG00000134744
ENSG00000166224	ENSG00000089154	ENSG00000132639	ENSG00000105655	ENSG00000120162	ENSG00000100297	ENSG00000006042
ENSG00000137513	ENSG00000083307	ENSG00000115541	ENSG00000172977	ENSG00000130055	ENSG00000153250	ENSG00000138768
ENSG00000163161	ENSG00000157823	ENSG00000197818	ENSG00000109101	ENSG00000170445	ENSG00000109189	ENSG00000198001
ENSG00000158411	ENSG00000138641	ENSG00000082701	ENSG00000100600	ENSG00000168818	ENSG00000100393	ENSG00000077254
ENSG00000141543	ENSG00000121031	ENSG00000088832	ENSG00000124772	ENSG00000165443	ENSG00000153481	ENSG00000182400
ENSG00000145794	ENSG00000124207	ENSG00000176476	ENSG00000165995	ENSG00000137474	ENSG00000185516	ENSG00000108510
ENSG00000161513	ENSG00000106524	ENSG00000153015	ENSG00000170854	ENSG00000111987	ENSG00000122954	ENSG00000130383
ENSG00000146733	ENSG00000139684	ENSG00000063854	ENSG00000186712	ENSG00000102189	ENSG00000184182	ENSG00000138757
ENSG00000125037	ENSG00000121749	ENSG00000089693	ENSG00000185721	ENSG00000110107	ENSG00000100823	ENSG00000139131
ENSG00000184840	ENSG00000084092	ENSG00000114331	ENSG00000076242	ENSG00000102743	ENSG00000164032	ENSG00000196235
ENSG00000182087	ENSG00000068654	ENSG00000105643	ENSG00000115561	ENSG00000106105	ENSG00000062650	ENSG00000169764
ENSG00000166974	ENSG00000143748	ENSG00000148200	ENSG00000033050	ENSG00000135624	ENSG00000156110	ENSG00000142507
ENSG00000118046	ENSG00000196449	ENSG00000164347	ENSG00000164172	ENSG00000166228	ENSG00000197872	ENSG00000130702
ENSG00000106443	ENSG00000166925	ENSG00000169925	ENSG00000111667	ENSG00000108091	ENSG00000099956	ENSG00000181449
ENSG00000104412	ENSG00000114200	ENSG00000160201	ENSG00000178878	ENSG00000115685	ENSG00000165672	ENSG00000116957
ENSG00000074855	ENSG00000115468	ENSG00000084072	ENSG00000178184	ENSG00000140859	ENSG00000131795	ENSG00000165960
ENSG00000130313	ENSG00000088930	ENSG00000168066	ENSG00000105135	ENSG00000162401	ENSG00000076650	ENSG00000164609
ENSG00000096150	ENSG00000136891	ENSG00000071189	ENSG00000130749	ENSG00000117054	ENSG00000108946	ENSG00000198807
ENSG00000065665	ENSG00000166192	ENSG00000129667	ENSG00000089094	ENSG00000003756	ENSG00000198721	ENSG00000198746
ENSG00000184983	ENSG00000119718	ENSG00000111142	ENSG00000115159	ENSG00000102312	ENSG00000163002	ENSG00000183049
ENSG00000057608	ENSG00000157890	ENSG00000126457	ENSG00000115825	ENSG00000135829	ENSG00000143499	ENSG00000107669
ENSG00000133997	ENSG00000080189	ENSG00000110063	ENSG00000156875	ENSG00000138279	ENSG00000163950	ENSG00000074695
ENSG00000165406	ENSG00000114166	ENSG00000106617	ENSG00000142657	ENSG00000104626	ENSG00000132970	ENSG00000112339
ENSG00000146267	ENSG00000112474	ENSG00000178913	ENSG00000100116	ENSG00000103356	ENSG00000159792	ENSG00000172061
ENSG00000181029	ENSG00000135476	ENSG00000120616	ENSG00000139352	ENSG00000182569	ENSG00000087460	ENSG00000082566
ENSG00000140521	ENSG00000141385	ENSG00000120008	ENSG00000100330	ENSG00000138326	ENSG00000105983	ENSG00000196177
ENSG00000099964	ENSG00000137770	ENSG00000171100	ENSG00000084733	ENSG00000133316	ENSG00000101210	ENSG00000149091
ENSG00000088038	ENSG00000144591	ENSG00000176890	ENSG00000144048	ENSG00000004961	ENSG00000184185	ENSG00000087510
ENSG00000093000	ENSG00000072310	ENSG00000132300	ENSG00000112685	ENSG00000167670	ENSG00000119950	ENSG00000145826
ENSG00000156261	ENSG00000074047	ENSG00000070061	ENSG00000104637	ENSG00000120705	ENSG00000182220	ENSG00000171135
ENSG00000128923	ENSG00000065328	ENSG00000110172	ENSG00000114126	ENSG00000127884	ENSG00000163623	ENSG00000123094
ENSG00000131051	ENSG00000108055	ENSG00000036257	ENSG00000103546	ENSG00000170312	ENSG00000185651	ENSG00000009780
ENSG00000138663	ENSG00000108443	ENSG00000090857	ENSG00000167325	ENSG00000105372	ENSG00000140933	ENSG00000047578
ENSG00000156011	ENSG00000167475	ENSG00000112379	ENSG00000196413	ENSG00000166598	ENSG00000110880	ENSG00000140829
ENSG00000140443	ENSG00000198677	ENSG00000136371	ENSG00000135404	ENSG00000083845	ENSG00000141388	ENSG00000175137
ENSG00000138744	ENSG00000116649	ENSG00000165704	ENSG00000105364	ENSG00000086061	ENSG00000171234	ENSG00000198558
ENSG00000182544	ENSG00000175093	ENSG00000198478	ENSG00000124541	ENSG00000120438	ENSG00000144061	ENSG00000101152
ENSG00000173376	ENSG00000087299	ENSG00000175287	ENSG00000059573	ENSG00000148225	ENSG00000100138	ENSG00000101444
ENSG00000198910	ENSG00000162300	ENSG00000180828	ENSG00000188818	ENSG00000095951	ENSG00000113851	ENSG00000149571
ENSG00000165898	ENSG00000103642	ENSG00000135821	ENSG00000175166	ENSG00000123201	ENSG00000138381	ENSG00000099940
ENSG00000130165	ENSG00000086475	ENSG00000169862	ENSG00000106348	ENSG00000072134	ENSG00000100983	ENSG00000136783
ENSG00000148943	ENSG00000144959	ENSG00000124641	ENSG00000183207	ENSG00000187540	ENSG00000104524	ENSG00000155313
ENSG00000170836	ENSG00000131381	ENSG00000160193	ENSG00000116151	ENSG00000146828	ENSG00000106038	ENSG00000111405
ENSG00000121316	ENSG00000163428	ENSG00000111530	ENSG00000084652	ENSG00000155438	ENSG00000075151	ENSG00000132341
ENSG00000163697	ENSG00000197402	ENSG00000113719	ENSG00000188677	ENSG00000128059	ENSG00000109806	ENSG00000125352
ENSG00000104343	ENSG00000182827	ENSG00000153107	ENSG00000166226	ENSG00000131375	ENSG00000077549	ENSG00000152315
ENSG00000178252	ENSG00000113569	ENSG00000116353	ENSG00000111674	ENSG00000197006	ENSG00000140043	ENSG00000164933
ENSG00000010017	ENSG00000095794	ENSG00000134049	ENSG00000141367	ENSG00000168724	ENSG00000136824	ENSG00000115484
ENSG00000106263	ENSG00000113456	ENSG00000115657	ENSG00000103335	ENSG00000143774	ENSG00000185973	ENSG00000189308
ENSG00000140612	ENSG00000133019	ENSG00000157426	ENSG00000135845	ENSG00000102144	ENSG00000071564	ENSG00000166260
ENSG00000070423	ENSG00000118939	ENSG00000198825	ENSG00000198408	ENSG00000166922	ENSG00000063438	ENSG00000189007
ENSG00000197563	ENSG00000180917	ENSG00000071082	ENSG00000149136	ENSG00000105245	ENSG00000157593	ENSG00000120697
ENSG00000091164	ENSG000000127412	ENSG00000122545	ENSG00000112592	ENSG00000079134	ENSG00000043591	ENSG00000167740
ENSG00000167513	ENSG00000004487	ENSG00000105401	ENSG00000186999	ENSG00000176658	ENSG00000163104	ENSG00000083642
ENSG00000175193	ENSG00000162923	ENSG00000138430	ENSG00000114346	ENSG00000151322	ENSG00000136758	ENSG00000115216
ENSG00000198862	ENSG00000175575	ENSG00000149485	ENSG00000187323	ENSG00000143476	ENSG00000152953	ENSG00000101577
ENSG00000132424	ENSG00000171530	ENSG00000002745	ENSG00000198931	ENSG00000080371	ENSG00000110046	ENSG00000148688
ENSG00000105341	ENSG00000099904	ENSG00000087095	ENSG00000145730	ENSG00000101890	ENSG00000133835	ENSG00000114316
ENSG00000006451	ENSG00000165914	ENSG00000058600	ENSG00000197121	ENSG00000152684	ENSG00000100523	ENSG00000152234
ENSG00000115350	ENSG00000151148	ENSG00000085999	ENSG00000010072	ENSG00000100219	ENSG00000070669	ENSG00000133812
ENSG00000101096	ENSG00000165186	ENSG00000104142	ENSG00000176009	ENSG00000086205	ENSG00000114423	ENSG00000131781
ENSG00000103671	ENSG00000172613	ENSG00000138083	ENSG00000187210	ENSG00000122378	ENSG00000156052	ENSG00000099899
ENSG00000154059	ENSG00000157985	ENSG00000174738	ENSG00000170004	ENSG00000085276	ENSG00000164506	ENSG00000143256
ENSG00000103051	ENSG00000085840	ENSG00000141127	ENSG00000176165	ENSG00000154945	ENSG00000113460	ENSG00000103342
ENSG00000100601	ENSG00000077235	ENSG00000185414	ENSG00000181789	ENSG00000137877	ENSG00000077097	ENSG00000112208
ENSG00000159650	ENSG00000118971	ENSG00000165304	ENSG00000158987	ENSG00000164764	ENSG00000106470	ENSG00000112237
ENSG00000116690	ENSG00000120925	ENSG00000109917	ENSG00000172238	ENSG00000079277	ENSG00000163357	ENSG00000160007
ENSG00000156113	ENSG00000100226	ENSG00000130396	ENSG00000049860	ENSG00000090694	ENSG00000100258	ENSG00000153922
ENSG000000015133	ENSG00000110583	ENSG00000075785	ENSG00000049656	ENSG00000117984	ENSG00000068796	ENSG00000167004
ENSG00000136463	ENSG00000136319	ENSG00000105993	ENSG00000137330	ENSG00000141646	ENSG00000166965	ENSG00000069329
ENSG00000156873	ENSG00000134001	ENSG00000165609	ENSG00000188869	ENSG00000143621	ENSG00000165417	ENSG00000148154
						ENSG00000131626
						ENSG00000100023
						ENSG00000183098
						ENSG00000126016
						ENSG00000162066

ENSG00000064313	ENSG00000184634	ENSG00000088812	ENSG00000110025	ENSG00000140519	ENSG00000197816	ENSG00000085231
ENSG00000134815	ENSG00000153933	ENSG00000065978	ENSG00000117222	ENSG00000164100	ENSG00000185681	ENSG00000110060
ENSG00000048544	ENSG00000104879	ENSG00000108590	ENSG00000073910	ENSG00000102786	ENSG00000162971	ENSG00000135334
ENSG00000143314	ENSG00000107815	ENSG00000196911	ENSG00000172663	ENSG00000107147	ENSG00000177042	ENSG00000134146
ENSG00000108582	ENSG00000106638	ENSG00000179104	ENSG00000039650	ENSG00000007062	ENSG00000164818	ENSG00000023909
ENSG00000140400	ENSG00000157916	ENSG00000132436	ENSG00000013375	ENSG00000184545	ENSG00000188659	ENSG00000137845
ENSG00000178381	ENSG00000125351	ENSG00000165630	ENSG00000112489	ENSG00000100262	ENSG00000197746	ENSG00000187109
ENSG00000140463	ENSG00000041982	ENSG00000121957	ENSG00000183760	ENSG00000131732	ENSG00000105792	ENSG00000115652
ENSG00000132388	ENSG00000139549	ENSG00000006715	ENSG00000139697	ENSG00000114541	ENSG00000160188	ENSG00000165195
ENSG00000077782	ENSG00000113558	ENSG00000173011	ENSG00000140995	ENSG00000111581	ENSG00000060138	ENSG00000152348
ENSG00000198728	ENSG00000129559	ENSG00000152942	ENSG00000057149	ENSG00000075415	ENSG00000088986	ENSG00000084774
ENSG00000137343	ENSG00000122687	ENSG00000141485	ENSG00000114204	ENSG00000146731	ENSG00000167578	ENSG00000198756
ENSG00000160209	ENSG00000131323	ENSG00000188909	ENSG00000160294	ENSG00000029725	ENSG00000165868	ENSG00000055130
ENSG00000156976	ENSG00000100266	ENSG00000135473	ENSG00000039654	ENSG00000109472	ENSG00000120262	ENSG00000099795
ENSG00000125107	ENSG00000120948	ENSG00000175985	ENSG00000160325	ENSG00000198183	ENSG00000139537	ENSG00000170522
ENSG00000108439	ENSG00000131238	ENSG00000138193	ENSG00000101350	ENSG00000196975	ENSG00000178125	ENSG00000063761
ENSG00000146109	ENSG00000129315	ENSG00000148835	ENSG00000147416	ENSG00000163060	ENSG00000163060	ENSG00000173599
ENSG00000134899	ENSG00000070081	ENSG00000119616	ENSG00000197890	ENSG00000178934	ENSG00000163576	ENSG00000175390
ENSG00000174851	ENSG00000159086	ENSG00000165283	ENSG00000155052	ENSG00000178104	ENSG00000155530	ENSG00000155961
ENSG00000135776	ENSG00000189338	ENSG00000090686	ENSG00000175467	ENSG00000178053	ENSG00000100565	ENSG00000112110
ENSG00000129596	ENSG00000166889	ENSG00000055813	ENSG00000104299	ENSG00000177664	ENSG00000157227	ENSG00000096401
ENSG00000031544	ENSG00000092108	ENSG00000181982	ENSG00000070010	ENSG00000167553	ENSG00000135338	ENSG00000112658
ENSG00000185088	ENSG00000172007	ENSG00000188157	ENSG00000090621	ENSG00000164111	ENSG00000164323	ENSG00000122591
ENSG00000100592	ENSG00000174276	ENSG00000125820	ENSG00000175344	ENSG00000163191	ENSG00000139496	ENSG00000165416
ENSG00000127554	ENSG00000100030	ENSG00000109814	ENSG00000137335	ENSG00000152076	ENSG00000079974	ENSG00000134748
ENSG00000108829	ENSG00000115241	ENSG00000164327	ENSG00000120800	ENSG00000137285	ENSG00000135702	ENSG00000163491
ENSG00000132207	ENSG00000169021	ENSG00000167693	ENSG00000138246	ENSG00000133665	ENSG00000116863	ENSG00000039319
ENSG00000068793	ENSG00000102003	ENSG00000061676	ENSG00000115234	ENSG00000127952	ENSG00000120647	ENSG00000198833
ENSG00000138433	ENSG00000145545	ENSG00000176102	ENSG00000162104	ENSG00000124688	ENSG00000161996	ENSG00000125814
ENSG00000160220	ENSG00000171152	ENSG00000174227	ENSG00000136100	ENSG00000117477	ENSG00000138032	ENSG00000130720
ENSG00000112159	ENSG00000077044	ENSG00000101138	ENSG00000147509	ENSG00000117266	ENSG00000196562	ENSG00000124614
ENSG00000175745	ENSG00000103319	ENSG00000160271	ENSG00000141556	ENSG00000114353	ENSG00000132321	ENSG00000145088
ENSG00000167881	ENSG00000161204	ENSG00000155189	ENSG00000140015	ENSG00000112964	ENSG00000135108	ENSG00000131473
ENSG00000136643	ENSG00000131236	ENSG00000108424	ENSG00000095564	ENSG00000106012	ENSG00000163728	ENSG00000101890
ENSG00000123297	ENSG00000182601	ENSG00000072501	ENSG00000104375	ENSG00000105479	ENSG00000100218	ENSG00000172680
ENSG00000136478	ENSG00000078140	ENSG00000151240	ENSG00000125505	ENSG00000104833	ENSG00000109680	ENSG00000160505
ENSG00000072756	ENSG00000121579	ENSG00000197905	ENSG00000007372	ENSG00000104237	ENSG00000139714	ENSG00000160629
ENSG00000109670	ENSG00000188167	ENSG00000136854	ENSG00000148175	ENSG00000096238	ENSG00000102531	ENSG00000118482
ENSG00000040341	ENSG00000089220	ENSG00000140374	ENSG00000115275	ENSG00000092820	ENSG00000138669	ENSG00000099769
ENSG00000155878	ENSG00000131013	ENSG00000136868	ENSG00000177058	ENSG00000075142	ENSG00000134376	ENSG00000072110
ENSG00000183508	ENSG00000159377	ENSG00000171634	ENSG00000166321	ENSG00000064199	ENSG00000139154	ENSG00000164344
ENSG00000162298	ENSG00000155111	ENSG00000103544	ENSG00000133731	ENSG0000014216	ENSG00000198089	ENSG00000132139
ENSG00000114942	ENSG00000151025	ENSG00000131748	ENSG00000132694	ENSG00000184874	ENSG00000164627	ENSG00000137808
ENSG00000109654	ENSG00000196656	ENSG00000168876	ENSG00000198026	ENSG00000089472	ENSG00000198003	ENSG00000105771
ENSG00000182333	ENSG00000112282	ENSG00000113643	ENSG00000100316	ENSG00000141519	ENSG00000132275	ENSG00000125409
ENSG00000078403	ENSG00000124208	ENSG00000135940	ENSG00000004139	ENSG00000169126	ENSG00000048342	ENSG00000103160
ENSG00000154975	ENSG00000173171	ENSG00000133110	ENSG00000106628	ENSG00000054267	ENSG00000151704	ENSG00000135387
ENSG00000109163	ENSG00000104695	ENSG00000142910	ENSG00000176986	ENSG00000104450	ENSG00000117450	ENSG00000181322
ENSG00000105607	ENSG00000133243	ENSG00000104762	ENSG00000143105	ENSG00000163808	ENSG00000166445	ENSG00000139620
ENSG00000114982	ENSG00000173366	ENSG00000159461	ENSG00000131504	ENSG00000139323	ENSG00000100583	ENSG00000158486
ENSG00000074582	ENSG00000177180	ENSG00000148690	ENSG00000177628	ENSG00000156876	ENSG00000100271	ENSG00000112081
ENSG00000155304	ENSG00000129255	ENSG00000103241	ENSG00000162733	ENSG00000114395	ENSG00000132677	ENSG00000198498
ENSG00000168493	ENSG00000159131	ENSG00000114503	ENSG00000115211	ENSG00000109576	ENSG00000064668	ENSG00000100982
ENSG00000188244	ENSG00000154027	ENSG00000196636	ENSG00000160917	ENSG00000104872	ENSG00000115524	ENSG00000176410
ENSG00000147421	ENSG00000109775	ENSG00000163638	ENSG00000126062	ENSG00000078687	ENSG00000175175	ENSG00000175104
ENSG00000163468	ENSG00000130255	ENSG00000145414	ENSG00000004866	ENSG00000108825	ENSG00000115947	ENSG00000171160
ENSG00000183496	ENSG00000001084	ENSG00000145331	ENSG00000103274	ENSG00000084693	ENSG00000104320	ENSG00000122692
ENSG00000103769	ENSG00000102038	ENSG00000183617	ENSG00000178921	ENSG00000110321	ENSG00000166323	ENSG00000137691
ENSG00000104388	ENSG00000089169	ENSG00000067225	ENSG00000172915	ENSG00000155761	ENSG00000136878	ENSG00000148356
ENSG00000143158	ENSG00000166377	ENSG00000115641	ENSG00000133773	ENSG00000131981	ENSG00000172070	ENSG00000177225
ENSG00000101193	ENSG00000144908	ENSG00000143106	ENSG00000163848	ENSG00000133739	ENSG00000115548	ENSG00000145075
ENSG00000104325	ENSG00000060069	ENSG00000143303	ENSG00000182899	ENSG00000104313	ENSG00000133114	ENSG00000133710
ENSG00000179194	ENSG00000170289	ENSG00000089157	ENSG00000109736	ENSG00000198782	ENSG00000118307	ENSG00000104044
ENSG00000170734	ENSG00000121644	ENSG00000140319	ENSG00000101425	ENSG00000181045	ENSG00000165698	ENSG00000184260
ENSG00000147465	ENSG00000197535	ENSG00000125755	ENSG00000094880	ENSG00000153904	ENSG00000164651	ENSG00000110002
ENSG00000196132	ENSG00000181222	ENSG00000144406	ENSG00000149483	ENSG00000055955	ENSG00000180715	ENSG00000101624
ENSG00000079313	ENSG00000115268	ENSG00000101391	ENSG00000147403	ENSG00000162814	ENSG00000136319	ENSG00000001143
ENSG00000108296	ENSG00000169718	ENSG00000197329	ENSG00000135972	ENSG00000157856	ENSG00000169402	ENSG00000133110
ENSG00000174485	ENSG000000011454	ENSG00000145020	ENSG00000105223	ENSG00000138892	ENSG00000135778	ENSG00000198677
ENSG00000042088	ENSG00000066583	ENSG00000152689	ENSG00000115649	ENSG00000108046	ENSG00000167210	ENSG000000012048
ENSG00000033867	ENSG00000180660	ENSG00000115504	ENSG000000013503	ENSG00000110429	ENSG00000115137	ENSG00000080572
ENSG00000100417	ENSG00000114742	ENSG00000120318	ENSG00000163703	ENSG00000183773	ENSG00000164070	ENSG00000100285
ENSG00000133026	ENSG00000103150	ENSG00000163541	ENSG00000147471	ENSG00000101350	ENSG00000008086	ENSG00000155636
ENSG00000197885	ENSG00000145780	ENSG0000013275	ENSG00000119523	ENSG00000101350	ENSG00000109846	ENSG00000070961
ENSG00000140905	ENSG00000138135	ENSG00000125954	ENSG00000140575	ENSG00000092850	ENSG00000164619	ENSG00000001630
ENSG00000158825	ENSG00000074621	ENSG00000150768	ENSG00000105438	ENSG00000182247	ENSG00000176273	ENSG00000119125
ENSG00000141279	ENSG00000118200	ENSG00000177602	ENSG00000099204	ENSG00000165886	ENSG00000151849	ENSG00000186063
ENSG00000163636	ENSG00000180875	ENSG00000129521	ENSG00000174165	ENSG00000164414	ENSG00000170502	ENSG00000159307
ENSG00000084110	ENSG00000169427	ENSG00000122008	ENSG00000196642	ENSG00000105982	ENSG00000114859	ENSG00000151014
ENSG00000134909	ENSG00000177084	ENSG00000103423	ENSG00000108587	ENSG00000123360	ENSG00000105640	ENSG00000185518
	ENSG00000188530					

ENSG00000118017	ENSG00000172062	ENSG00000142700	ENSG00000161328	ENSG00000166164
ENSG00000066279	ENSG00000167136	ENSG00000165282	ENSG00000178202	ENSG00000124275
ENSG00000166169	ENSG00000147677	ENSG00000135638	ENSG00000153575	ENSG00000198690
ENSG00000130706	ENSG00000112701	ENSG00000114686	ENSG00000135677	ENSG00000171055
ENSG00000121057	ENSG00000178607	ENSG00000174013	ENSG00000167118	ENSG00000128694
ENSG00000134717	ENSG00000106689	ENSG00000163939	ENSG00000142208	ENSG00000109062
ENSG00000113597	ENSG00000116198	ENSG00000143486	ENSG00000188329	ENSG00000005810
ENSG00000101421	ENSG00000136449	ENSG00000167770	ENSG00000127980	ENSG00000155957
ENSG00000155970	ENSG00000135315	ENSG00000123728	ENSG00000064999	ENSG00000100033
ENSG00000089050	ENSG00000174903	ENSG00000092020	ENSG00000165118	ENSG00000007392
ENSG00000124275	ENSG00000135720	ENSG00000107862	ENSG00000168016	ENSG00000149823
ENSG00000130772	ENSG00000188620	ENSG00000081721	ENSG00000185798	ENSG00000144579
ENSG00000138231	ENSG00000113722	ENSG00000067167	ENSG00000122705	ENSG00000152404
ENSG00000102580	ENSG00000165322	ENSG00000148459	ENSG00000126778	ENSG00000151247
ENSG00000175066	ENSG00000106268	ENSG00000115252	ENSG00000120265	ENSG00000091009
ENSG00000006712	ENSG00000113013	ENSG00000165905	ENSG00000170426	ENSG00000143761
ENSG00000118518	ENSG00000173145	ENSG00000138303	ENSG00000174327	ENSG00000130150
ENSG00000164074	ENSG00000100439	ENSG00000132541	ENSG00000141564	ENSG00000125691
ENSG00000135862	ENSG00000100664	ENSG00000115593	ENSG00000168738	ENSG00000131373
ENSG00000162390	ENSG00000180979	ENSG00000178695	ENSG00000145782	ENSG00000139719
ENSG00000065150	ENSG00000085721	ENSG00000126432	ENSG00000151849	ENSG00000109686
ENSG00000187773	ENSG00000168129	ENSG00000116096	ENSG00000116747	ENSG00000168591
ENSG00000112699	ENSG00000136875	ENSG00000160993	ENSG00000156172	ENSG00000111596
ENSG00000165650	ENSG00000102978	ENSG00000135049	ENSG00000188010	ENSG00000128050
ENSG00000110628	ENSG00000005206	ENSG00000117151	ENSG00000137343	ENSG00000138604
ENSG00000198162	ENSG00000170941	ENSG00000103042	ENSG00000165502	ENSG00000167258
ENSG00000161057	ENSG00000167632	ENSG00000129295	ENSG00000103494	ENSG00000153574
ENSG00000125743	ENSG00000160293	ENSG00000104808	ENSG00000132768	ENSG00000137955
ENSG00000054523	ENSG00000077721	ENSG00000167131	ENSG00000074319	ENSG00000107731
ENSG00000158006	ENSG00000163517	ENSG00000137171	ENSG00000175938	ENSG00000143815
ENSG00000120907	ENSG00000130227	ENSG00000120440	ENSG00000105426	ENSG00000140632
ENSG00000170820	ENSG00000100207	ENSG00000151715	ENSG00000119686	ENSG00000054356
ENSG00000176715	ENSG00000151079	ENSG00000082153	ENSG00000106211	ENSG00000163354
ENSG00000176201	ENSG00000153936	ENSG00000174996	ENSG00000068383	ENSG00000144136
ENSG00000185130	ENSG00000161202	ENSG00000152932	ENSG00000159063	ENSG00000120586
ENSG00000181830	ENSG00000116337	ENSG00000116001	ENSG00000115677	ENSG00000112062
ENSG00000108848	ENSG00000113282	ENSG00000196433	ENSG00000112333	ENSG00000101266
ENSG00000157445	ENSG00000162992	ENSG00000129071	ENSG00000156502	ENSG00000160211
ENSG00000089685	ENSG00000164114	ENSG00000143199	ENSG00000142230	ENSG00000100416
ENSG00000143079	ENSG00000163911	ENSG00000187908	ENSG00000088035	ENSG00000102444
ENSG00000140988	ENSG00000196365	ENSG00000116685	ENSG00000197976	ENSG00000171603
ENSG00000179580	ENSG00000113312	ENSG00000131100	ENSG00000163539	ENSG00000091127
ENSG00000120868	ENSG00000173811	ENSG00000162392	ENSG00000112498	ENSG00000169696
ENSG00000173546	ENSG00000185141	ENSG00000159409	ENSG00000059758	ENSG00000167646
ENSG00000186417	ENSG00000163624	ENSG00000160439	ENSG00000182010	ENSG00000086288
ENSG00000132911	ENSG00000197114	ENSG00000100364	ENSG00000089060	ENSG00000165506
ENSG00000182831	ENSG00000198929	ENSG00000116035	ENSG00000100284	ENSG00000157423
ENSG00000177971	ENSG00000165105	ENSG00000163785	ENSG00000185624	ENSG00000046651
ENSG00000164187	ENSG00000182786	ENSG00000110700	ENSG00000155666	ENSG00000156313
ENSG00000130985	ENSG00000128609	ENSG00000086102	ENSG00000149260	ENSG00000138190
ENSG00000180817	ENSG00000167701	ENSG00000148843	ENSG00000163808	ENSG00000175792
ENSG00000134109	ENSG00000003147	ENSG00000115919	ENSG00000100575	ENSG00000075568
ENSG00000031698	ENSG00000163877	ENSG00000100902	ENSG00000167393	ENSG00000073803
ENSG00000175595	ENSG00000073921	ENSG00000070785	ENSG00000023608	
ENSG00000185418	ENSG00000176887	ENSG00000163956	ENSG00000177000	
ENSG00000141627	ENSG00000155158	ENSG00000174576	ENSG00000136842	
ENSG00000196367	ENSG00000110066	ENSG00000101346	ENSG00000164860	
ENSG00000126005	ENSG00000138696	ENSG00000179761	ENSG00000028528	
ENSG00000085563	ENSG00000166275	ENSG00000163818	ENSG00000121989	
ENSG00000104687	ENSG00000058600	ENSG00000095596	ENSG00000177666	
ENSG00000060237	ENSG00000151576	ENSG00000101057	ENSG00000115514	
ENSG00000149292	ENSG00000175344	ENSG00000164118	ENSG00000107959	
ENSG00000100300	ENSG00000023318	ENSG00000170927	ENSG00000160179	
ENSG00000105865	ENSG00000067829	ENSG00000179889	ENSG00000183826	
ENSG00000144182	ENSG00000173826	ENSG00000177885	ENSG00000164244	
ENSG00000066279	ENSG00000091010	ENSG00000118194	ENSG00000083520	
ENSG00000169174	ENSG00000172469	ENSG00000044446	ENSG00000196151	
ENSG00000100324	ENSG00000105835	ENSG00000116663	ENSG00000161326	
ENSG00000127481	ENSG00000134321	ENSG00000171824	ENSG00000139239	
ENSG00000107130	ENSG00000076108	ENSG00000132676	ENSG00000069399	
ENSG00000004766	ENSG00000151575	ENSG00000129932	ENSG00000135392	
ENSG00000006625	ENSG00000133393	ENSG00000179598	ENSG00000105011	
ENSG00000110801	ENSG00000166333	ENSG00000168079	ENSG00000113712	
ENSG00000176261	ENSG00000004455	ENSG00000112498	ENSG00000141551	
ENSG00000171307	ENSG00000164898	ENSG00000008869	ENSG00000110514	
ENSG00000182473	ENSG00000105486	ENSG00000107902	ENSG00000153558	
ENSG00000172732	ENSG00000078177	ENSG00000101557	ENSG00000150760	
ENSG00000111110	ENSG00000137628	ENSG00000156869	ENSG00000146242	
ENSG00000123427	ENSG00000151790	ENSG00000080503	ENSG00000106976	

Table 10.2 Ensembl IDs of all genes found in the Ciliome database (www.sfu.ca/~leroux/ciliome_database.htm).

Example config file for Control-FREEC (for family 1 individual 1)

```

[general]
chrLenFile = hg19.len
window = 1000
step = 500
ploidy = 2
outputDir = /data/home/epmmee/FREEC/

GCcontentProfile = GC_profile_50kb.cnp

#sex=XY
breakPointType=4

gemMappabilityFile = /data/home/epmmee/FREEC/out100m2_hg19.gem
chrFiles = /data/home/epmmee/hg19 FASTA files/
samtools = /data/home/epmmee/samtools-0.1.19/samtools

maxThreads=6

breakPointThreshold=1.5
noisyData=TRUE
printNA=FALSE
BedGraphOutput=TRUE

[sample]

mateFile = /data/home/epmmee/FREEC/fam1_ind1.bam

inputFormat = 0
matesOrientation = FR

[control]

mateFile = /data/home/epmmee/FREEC/fam4_ind1.bam

inputFormat = 0
matesOrientation = FR

[BAF]
SNPfile = /data/home/epmmee/FREEC/hg19_snp137.SingleDiNucl.1based.txt
minimalCoveragePerPosition = 5

[target]

captureRegions = /data/home/epmmee/FREEC/truseq_exome_targeted_regions.hg19.bed.chr

```

9.5.5. Scripts used in PCD analyses

These scripts were used to extract suspected human PCD causal genes from public databases.

PCD_candidate_genes.py

```
WorkingDirectory = '/data/home/epmmee/SA_PCD_data/'
import sys
ciliome_file = open(WorkingDirectory + 'ciliome.csv', 'r')
output_file = open(WorkingDirectory + 'ciliome_genes.txt', 'w')
for line in ciliome_file:
    line = line.replace('\n', '')
    record = line.split(',')
    ensg_col = record[0]
    output_file.write(str(ensg_col) + '\n')
```

PCD_candidate_genes_location.py

```
WorkingDirectory = '/data/home/epmmee/SA_PCD_data/'
import sys
ciliome_mutation_file = open(WorkingDirectory +
'AD_NR_candidate_ciliome_mutations.txt', 'r')
output_file = open(WorkingDirectory +
'AD_NR_candidate_ciliome_genes_location.txt', 'w')
for line in ciliome_mutation_file:
    line = line.replace('\n', '')
    record = line.split('\t')
    location_col = record[1]
    location = location_col.split(':')
    mut_loc = location[1]
    output_file.write(str(mut_loc) + '\n')
```

9.5.6. Protein structure modelling

PDB data format

Columns	Data type	Contents
1	Record name	ATOM
2	Integer	Atom serial number
3	Atom	Atom name
4	Residue Name	Residue
5	Character	Residue Sequence no
6	Real(8.3)	Orthogonal coordinates for X in Angstroms
7	Real(8.3)	Orthogonal coordinates for Y in Angstroms
8	Real(8.3)	Orthogonal coordinates for Z in Angstroms
9	Real(6.2)	Occupancy
10	Real(6.2)	Temperature factor (Default = 0.0)

Table 10.3 Legend for the 10 columns of data in the PDB file (used in this thesis). These are used to model and view the structure of the protein (using RasMol).

MNS1

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-80	Alignment	1jadB_301	1-242	0.4844	Hydrolase
81-215	Alignment	3k29A_301	1-162	0.2239	Unknown function
216-400	Alignment	3dytA_303	1-355	0.1147	Transport protein
401-495	Alignment	1f5nA_301	1-570	0.2561	Signaling protein

Table 10.4 PDB Domains used for Ginzu Prediction when predicting MNS1p structure

Wild type sequence

MGSKRRNLSCSERHQKLVDENYCKKLHVQALKNVNSQIRNQMVQNENDNRVQRKQFLRLQLNEQFELDMEEAIQK
 AEENKRLKELQLKQEEKLAMELAKLKHESLKDEKMRQQVRENSIELRELEKKLKAAYMNKERAQAIAEKDAIKYE
 QMKRDAEIAKTMMEEHKRIIKEENAAEDKRNKAKAQYYLDLEKQLEEQEKKKQEAYEQLLKEKLMIDEIVRKIYE
 EDQLEKQQKLEKMNAMRRYIEEFQKEQALWRKKKREEEENNRKIEFANMQQQREEDRMAKVQENEEKRLQLQN
 ALTQKLEEMLRQREDLEQVRQELYQEEQAEIYKSKLKEEAEEKLRKQKEMKQDFEEQMALKELVLQAAKEEEENF
 RKTMLAKFAEDDRIELMNAQKQRMKQLEHRRRAVEKLIIEERRQQFLADKQRELEEWQLQQRRQGFINAIEEERLK
 LLKEHATNLLGYLPKGVFKKEDDIDLLGEEFRKVYQQRSEICEEK

Mutant-type sequence (p.M263T)

MGSKRRNLSCSERHQKLVDENYCKKLHVQALKNVNSQIRNQMVQNENDNRVQRKQFLRLQLNEQFELDMEEAIQK
 AEENKRLKELQLKQEEKLAMELAKLKHESLKDEKMRQQVRENSIELRELEKKLKAAYMNKERAQAIAEKDAIKYE
 QMKRDAEIAKTMMEEHKRIIKEENAAEDKRNKAKAQYYLDLEKQLEEQEKKKQEAYEQLLKEKLMIDEIVRKIYE
 EDQLEKQQKLEKMNAMRRYIEEFQKEQALWRKKKREEEENNRKIEFANMQQQREEDRMAKVQENEEKRLQLQN
 ALTQKLEEMLRQREDLEQVRQELYQEEQAEIYKSKLKEEAEEKLRKQKEMKQDFEEQMALKELVLQAAKEEEENF
 RKTMLAKFAEDDRIELMNAQKQRMKQLEHRRRAVEKLIIEERRQQFLADKQRELEEWQLQQRRQGFINAIEEERLK
 LLKEHATNLLGYLPKGVFKKEDDIDLLGEEFRKVYQQRSEICEEK

Wild type MNS1 PDB (at position 263) – Model 1

ATOM	4538	N	MET	263	-100.066	72.023	99.521	1.00	19.81
ATOM	4539	CA	MET	263	-101.464	72.427	99.613	1.00	23.32
ATOM	4540	C	MET	263	-102.267	71.444	100.452	1.00	22.66
ATOM	4541	O	MET	263	-103.115	71.866	101.239	1.00	23.92
ATOM	4542	CB	MET	263	-102.071	72.549	98.217	1.00	34.98
ATOM	4543	CG	MET	263	-101.540	73.718	97.399	1.00	34.98
ATOM	4544	SD	MET	263	-102.215	73.760	95.727	1.00	34.98
ATOM	4545	CE	MET	263	-103.922	74.187	96.062	1.00	34.98
ATOM	4546	H	MET	263	-99.685	71.813	98.610	1.00	23.77
ATOM	4547	HA	MET	263	-101.539	73.393	100.111	1.00	27.98
ATOM	4548	1HB	MET	263	-101.860	71.616	97.697	1.00	41.98
ATOM	4549	2HB	MET	263	-103.148	72.657	98.347	1.00	41.98
ATOM	4550	1HG	MET	263	-101.802	74.641	97.915	1.00	41.98
ATOM	4551	2HG	MET	263	-100.455	73.628	97.344	1.00	41.98
ATOM	4552	1HE	MET	263	-104.472	74.252	95.124	1.00	41.98
ATOM	4553	2HE	MET	263	-104.371	73.421	96.695	1.00	41.98
ATOM	4554	3HE	MET	263	-103.962	75.150	96.573	1.00	41.98

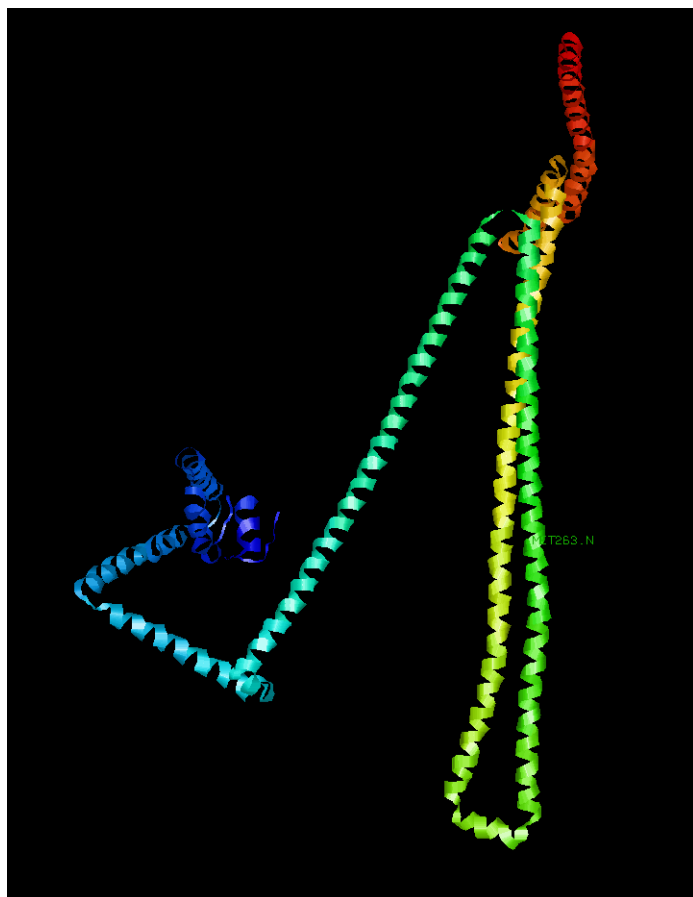
Mutant MNS1 PDB (at position 263) – Model 1

ATOM	4538	N	THR	263	-27.406	31.585	-82.879	1.00	30.33
ATOM	4539	CA	THR	263	-27.130	32.981	-83.177	1.00	26.19
ATOM	4540	C	THR	263	-26.123	33.108	-84.309	1.00	24.46
ATOM	4541	O	THR	263	-26.286	33.980	-85.161	1.00	23.54
ATOM	4542	CB	THR	263	-26.598	33.728	-81.940	1.00	39.29
ATOM	4543	OG1	THR	263	-27.587	33.707	-80.903	1.00	39.29
ATOM	4544	CG2	THR	263	-26.266	35.171	-82.291	1.00	39.29
ATOM	4545	H	THR	263	-27.164	31.237	-81.962	1.00	36.40
ATOM	4546	HG1	THR	263	-28.008	32.844	-80.881	1.00	47.14
ATOM	4547	HA	THR	263	-28.042	33.473	-83.517	1.00	31.43
ATOM	4548	HB	THR	263	-25.699	33.227	-81.583	1.00	47.14
ATOM	4549	1HG2	THR	263	-25.891	35.683	-81.404	1.00	47.14
ATOM	4550	2HG2	THR	263	-25.506	35.190	-83.071	1.00	47.14
ATOM	4551	3HG2	THR	263	-27.165	35.673	-82.647	1.00	47.14

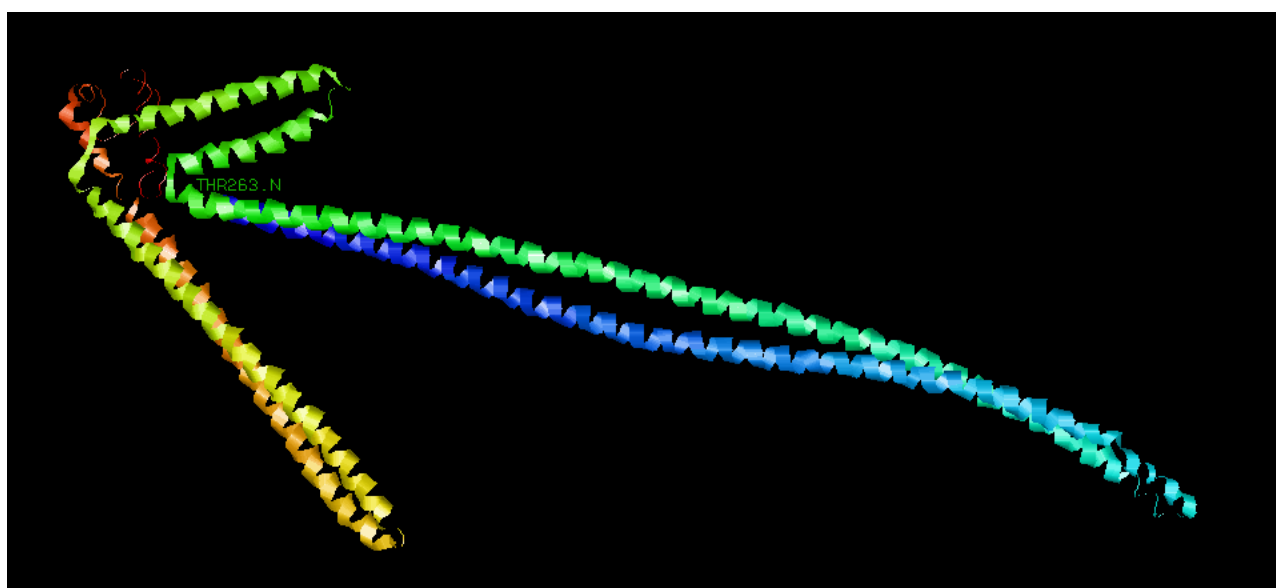
Other MNS1 Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

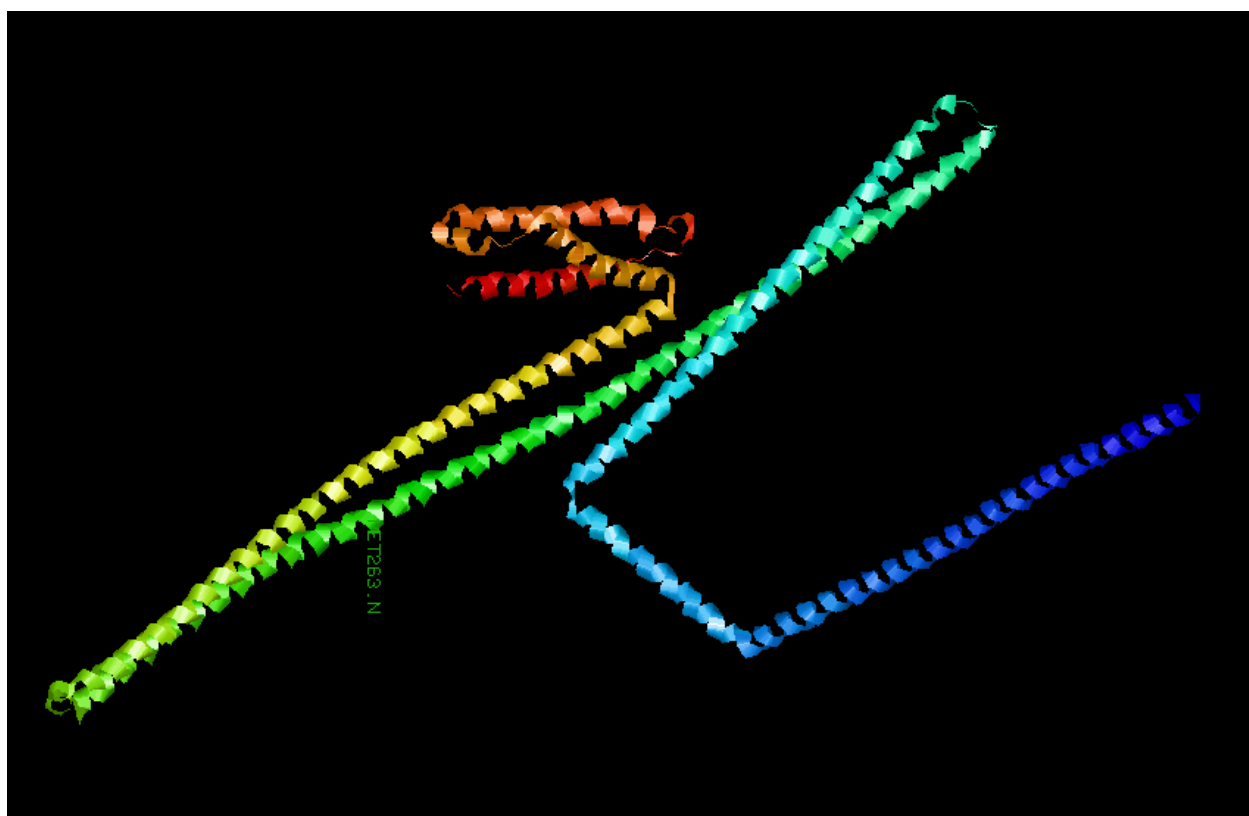
Model 2 – Wild Type



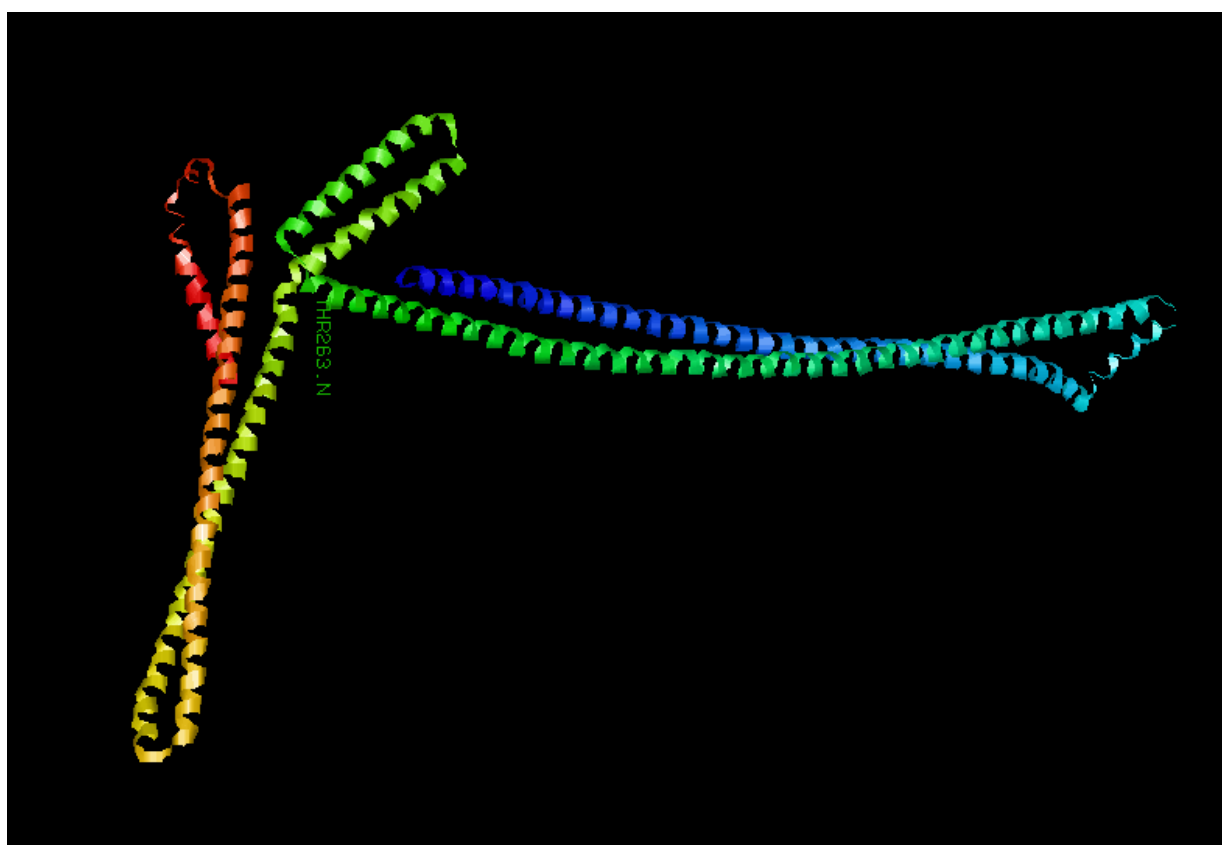
Model 2 – Mutant



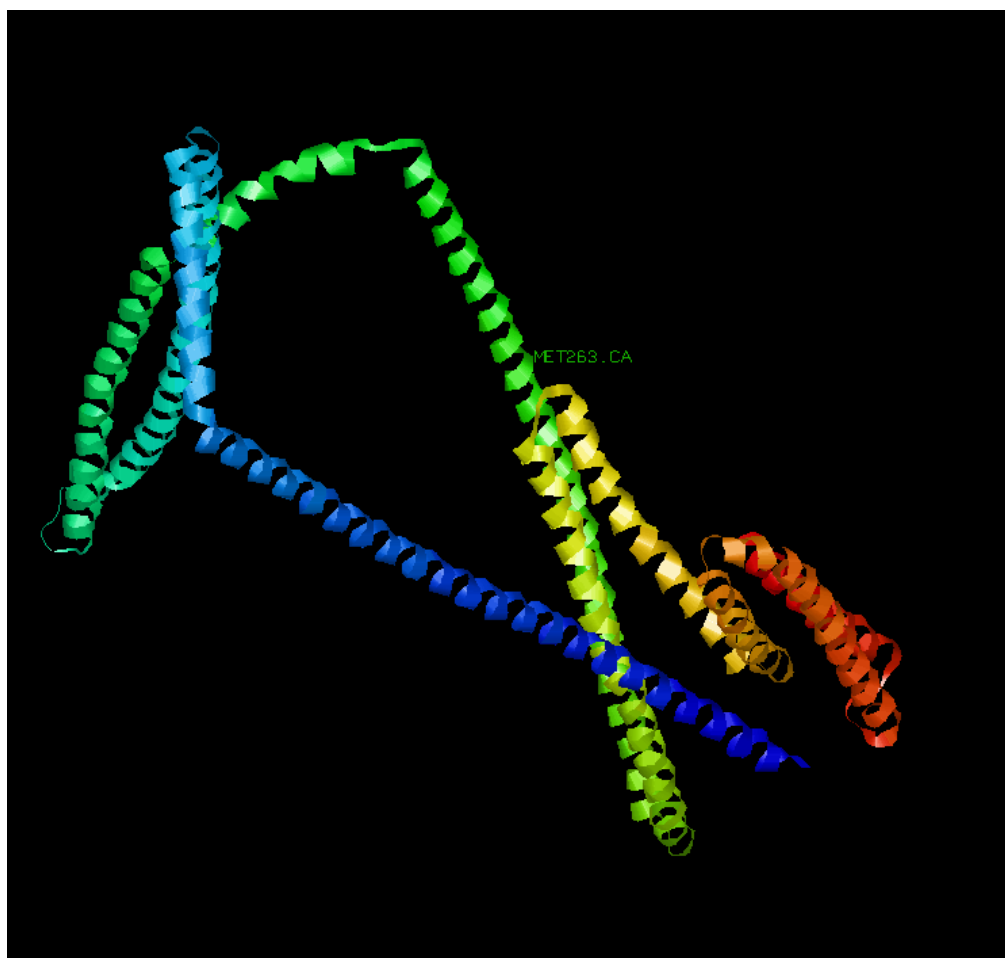
Model 3 – Wild Type



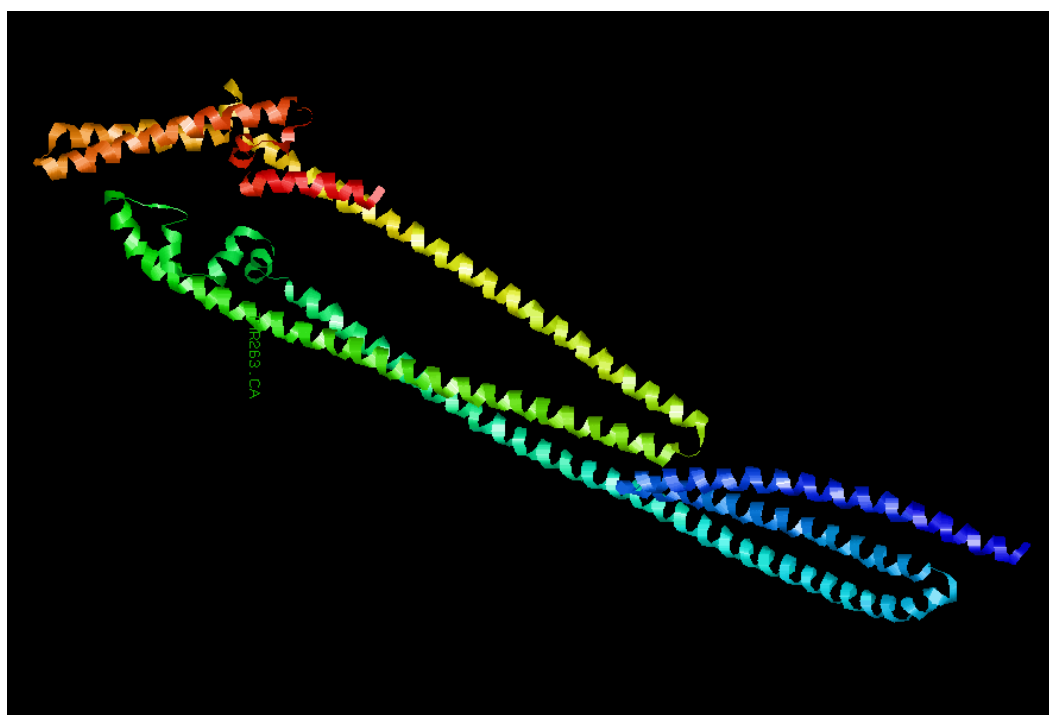
Model 3 – Mutant



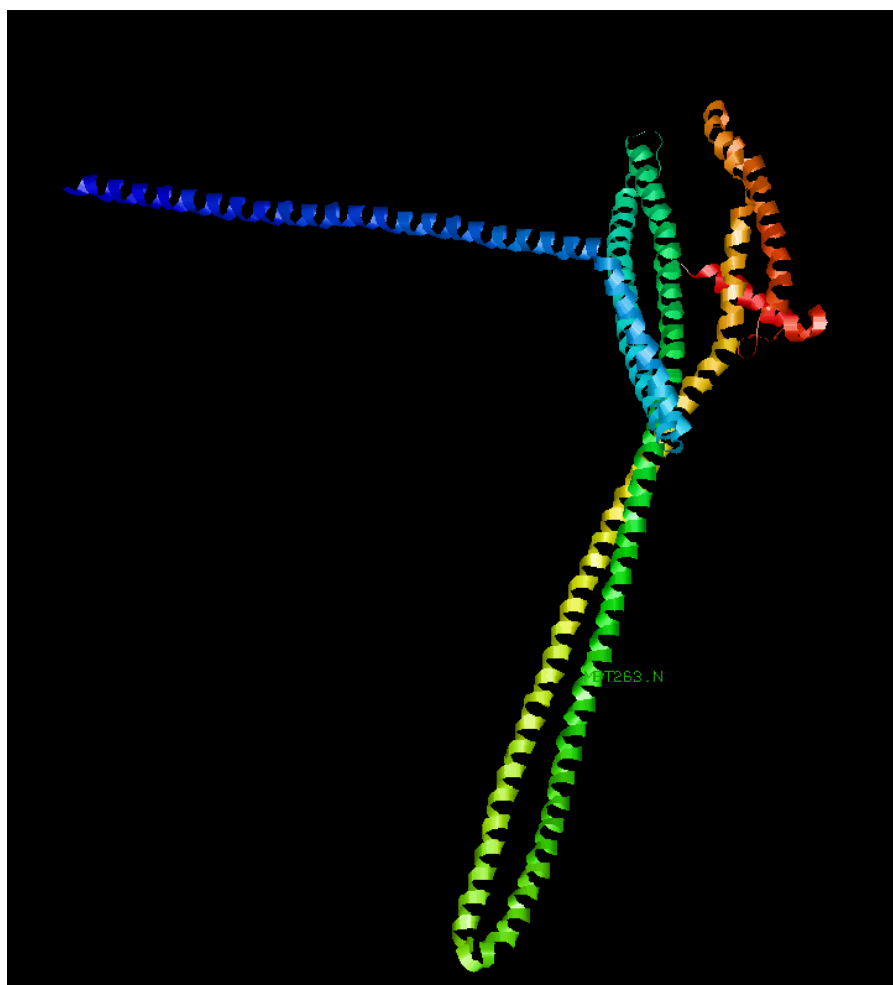
Model 4 - Wild Type



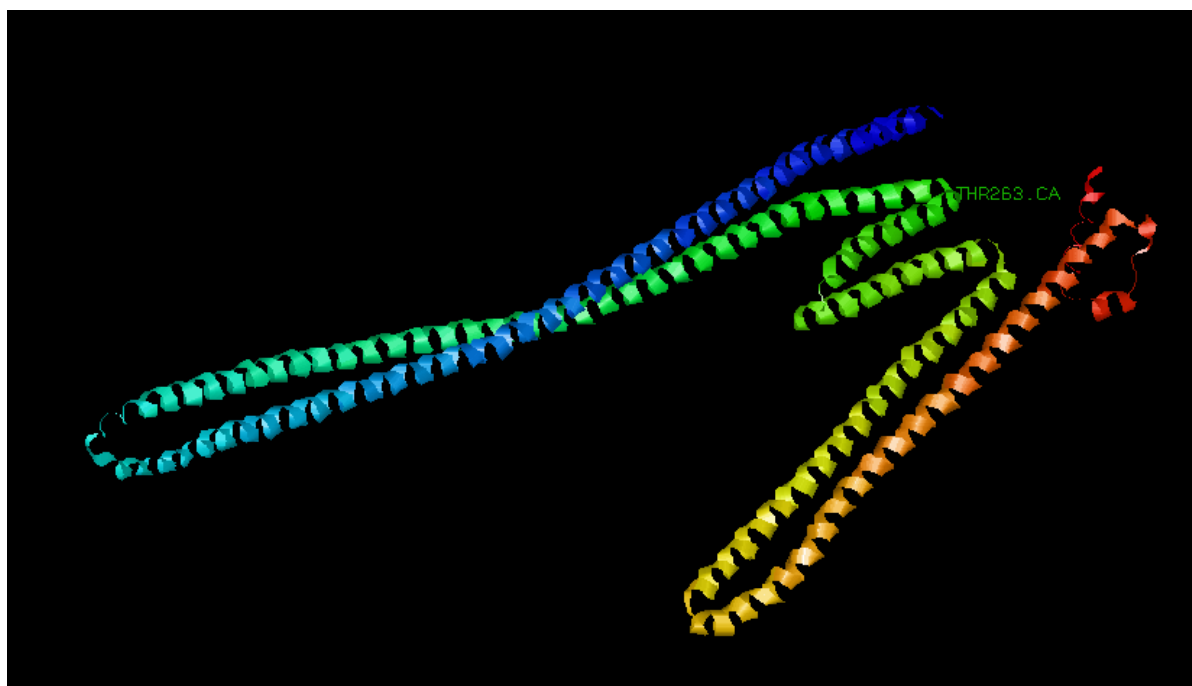
Model 4 - Mutant



Model 5 – Wild Type



Model 5 – Mutant



DNALI1

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-74	Alignment	4o8uA_101	1-227	0.3282	-
75-258	Alignment	3sogA_304	1-197	0.1808	Structural protein

Table 10.5 PDB Domains used for Ginzu Prediction when predicting DNALI1p structure

Wild type sequence

MIPPADSLKDYDTPVLVSRNTEKRSPKARLLKVSPQQPGPSGSAPQPPKTKLPSTPCVPDPTKQAEIILNAILPP
 REWVEDTQLWIIQQVSSTPSTRMDVVHLQEQDLKLQQRQARETGICPVRRELYSQCFDELIREVTINCAERGLLL
 LRVRDEIRMTIAAYQTLYESSVAFGMRKALQAEQKGSDMERKIAELETEKRDLERQVNEQKAKCEATEKRESERR
 QVEEKKHNEEIQFLKSTNQQLKAQLEGIIAPKK

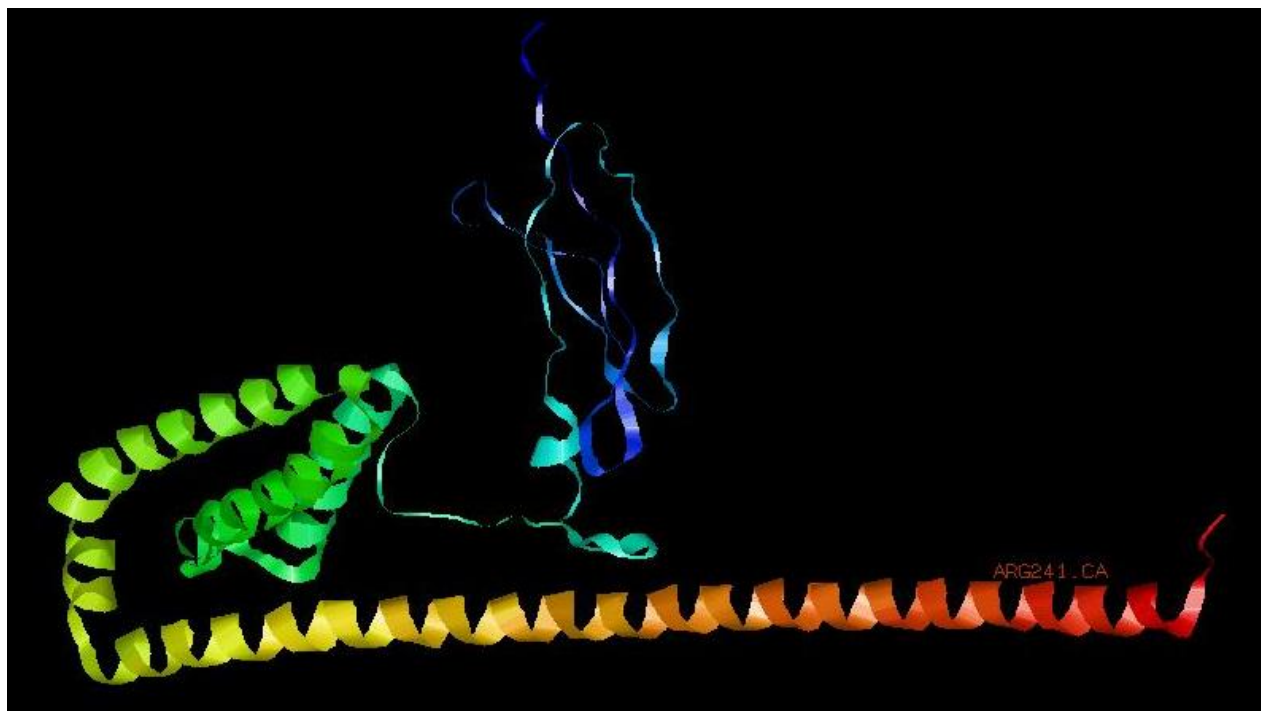
Mutant-type sequence (p.R263*)

MIPPADSLKDYDTPVLVSRNTEKRSPKARLLKVSPQQPGPSGSAPQPPKTKLPSTPCVPDPTKQAEIILNAILPP
 REWVEDTQLWIIQQVSSTPSTRMDVVHLQEQDLKLQQRQARETGICPVRRELYSQCFDELIREVTINCAERGLLL
 LRVRDEIRMTIAAYQTLYESSVAFGMRKALQAEQKGSDMERKIAELETEKRDLERQVNEQKAKCEATEKRESERR
 QVEEKKHNEEIQFLK

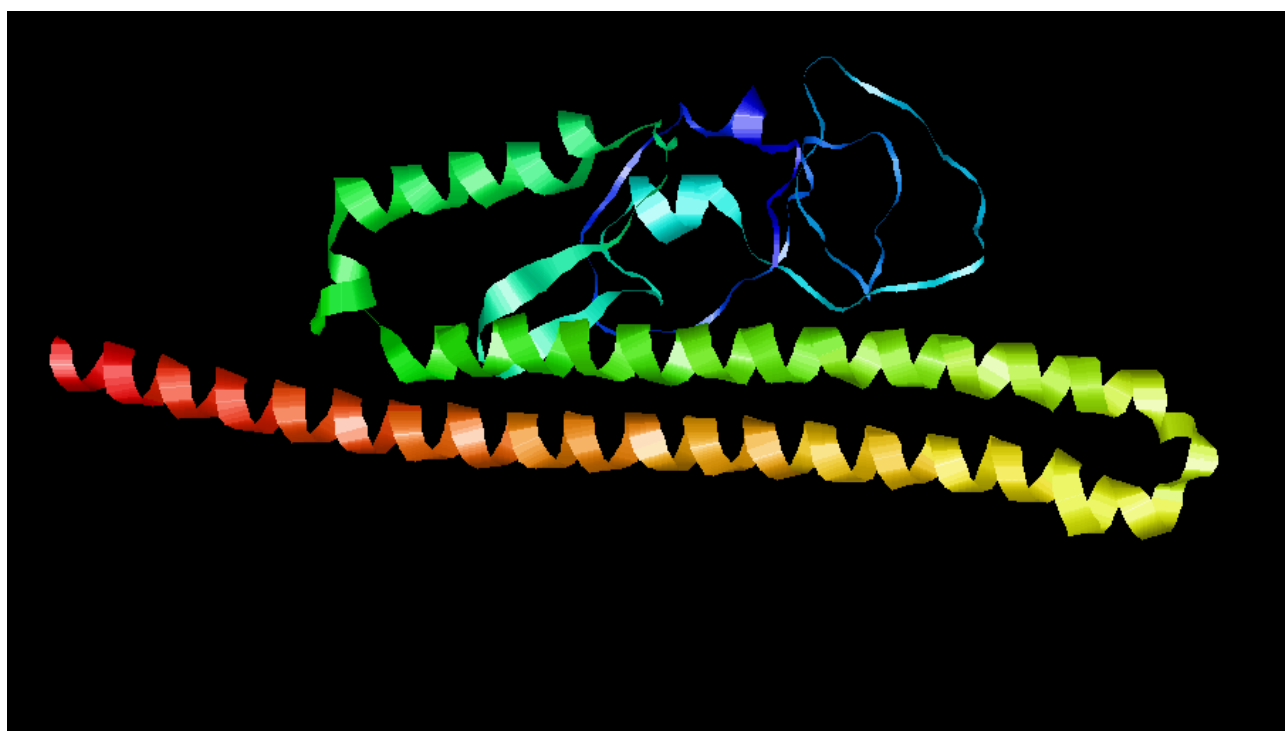
Other DNALI1 Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

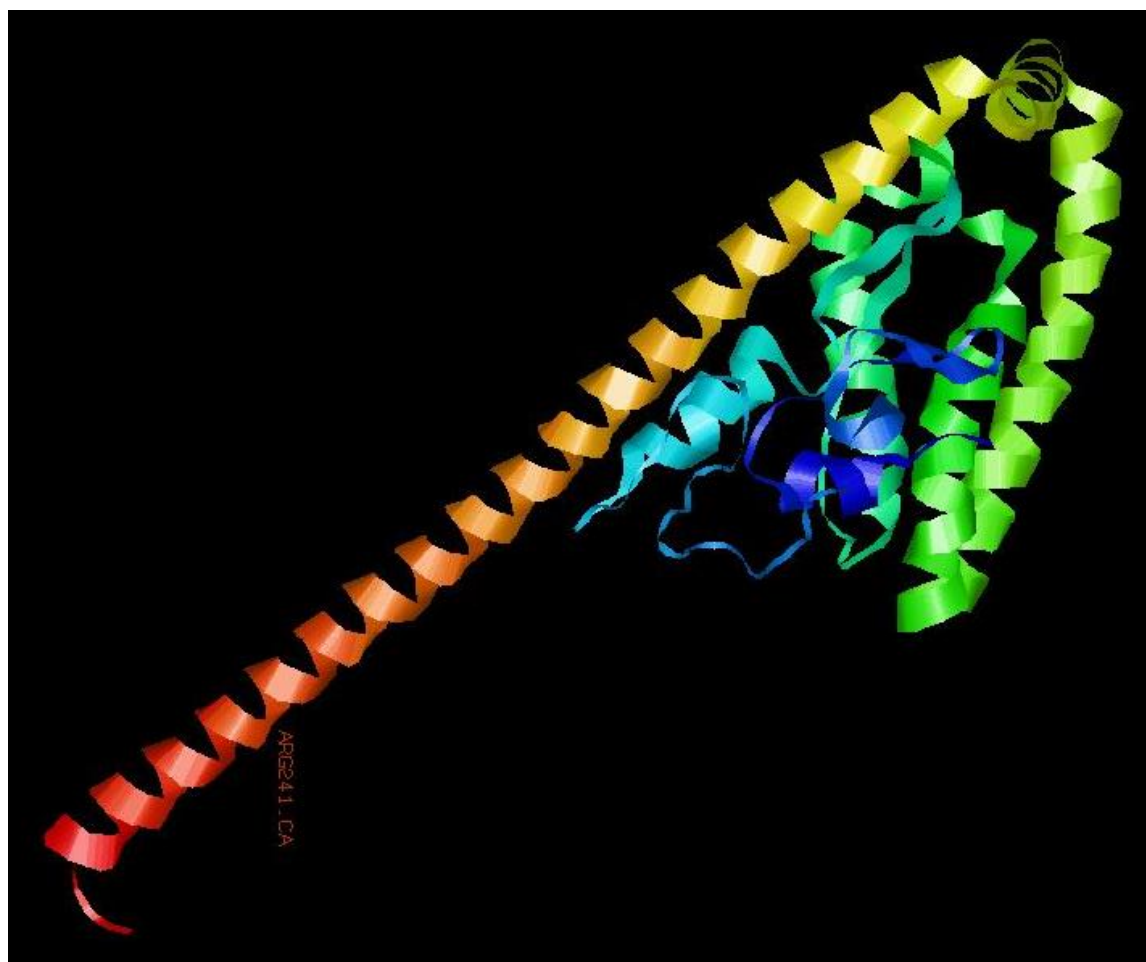
Model 2 – Wild Type



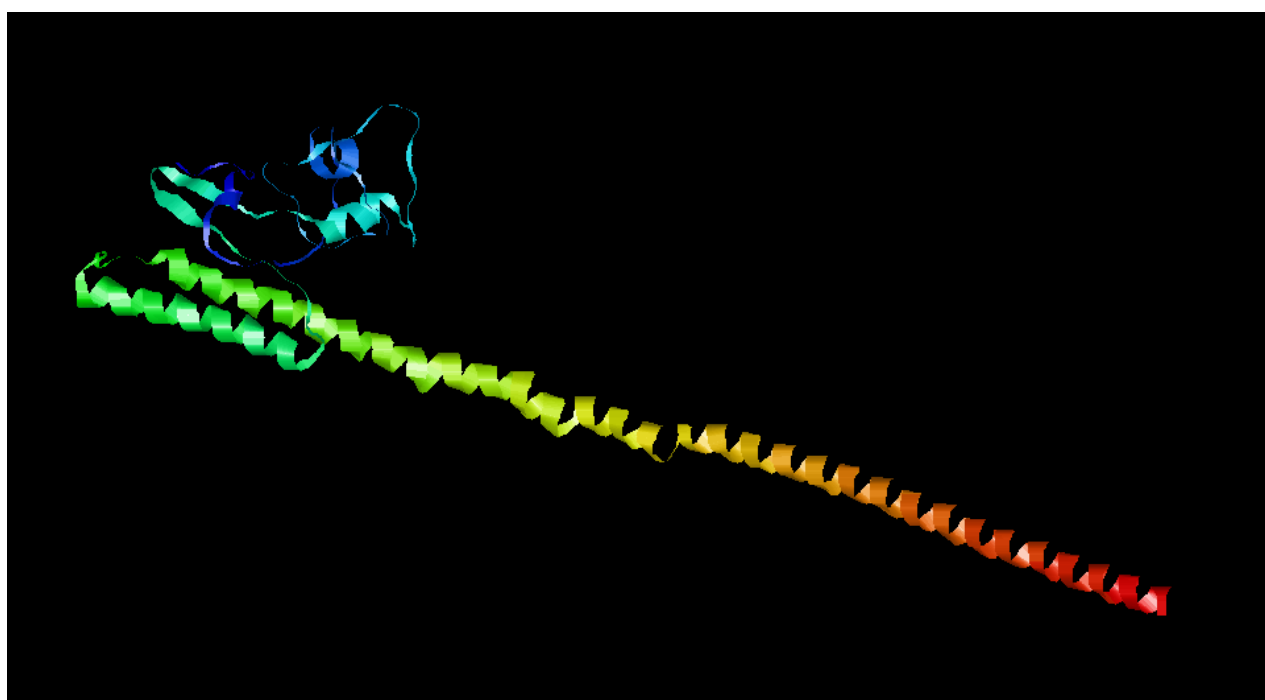
Model 2 – Mutant



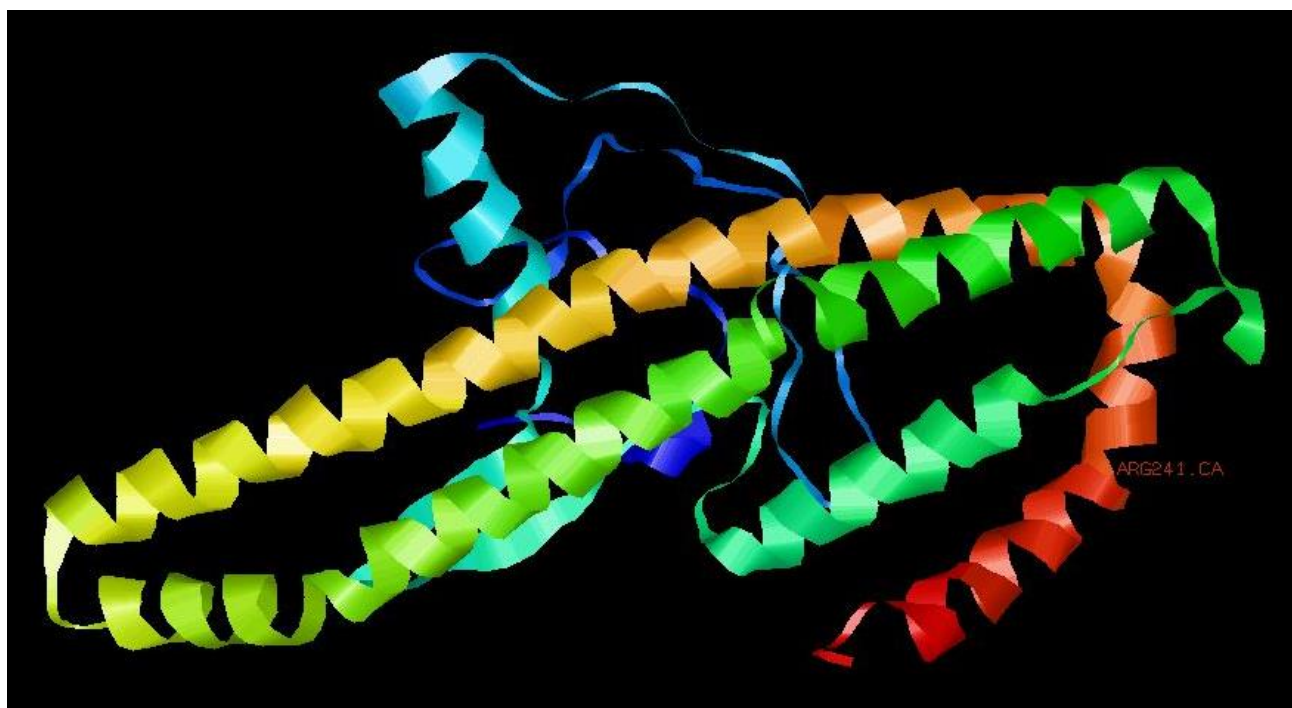
Model 3 – Wild Type



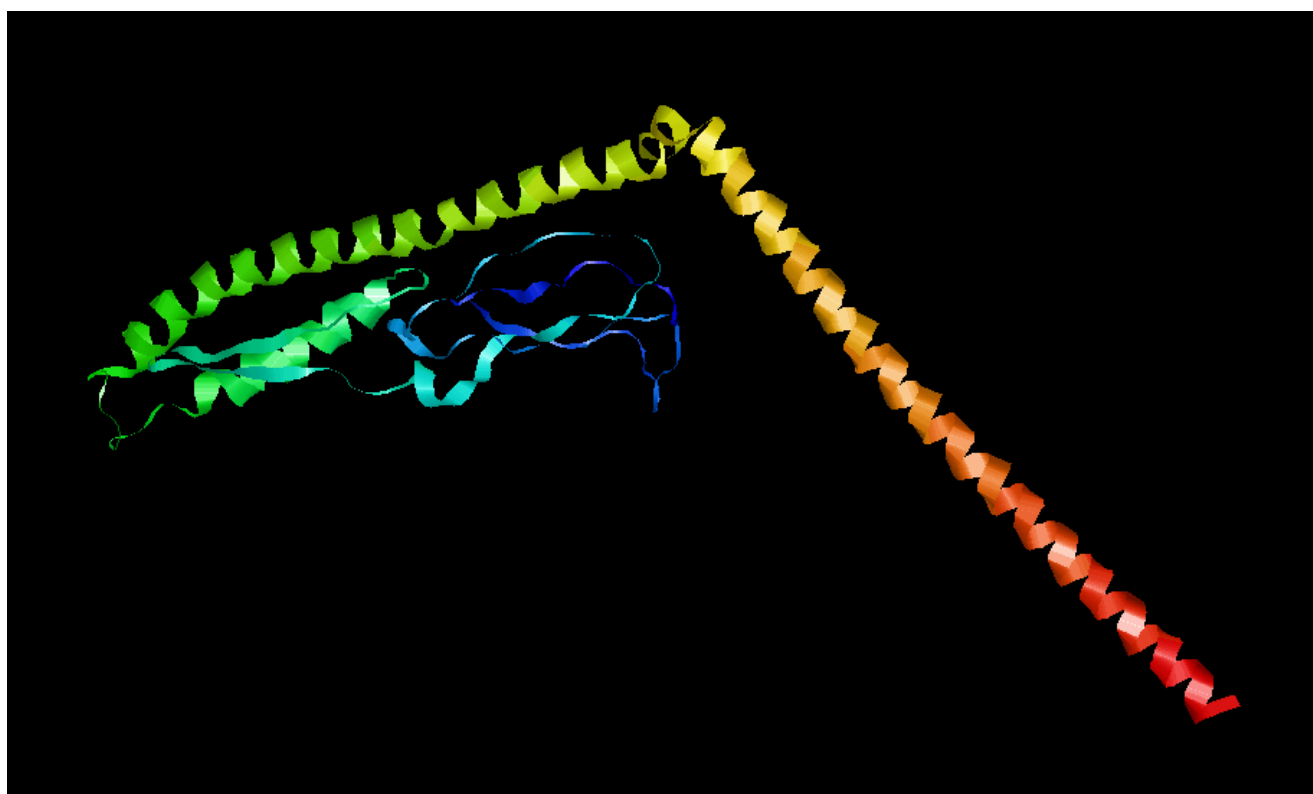
Model 3 – Mutant



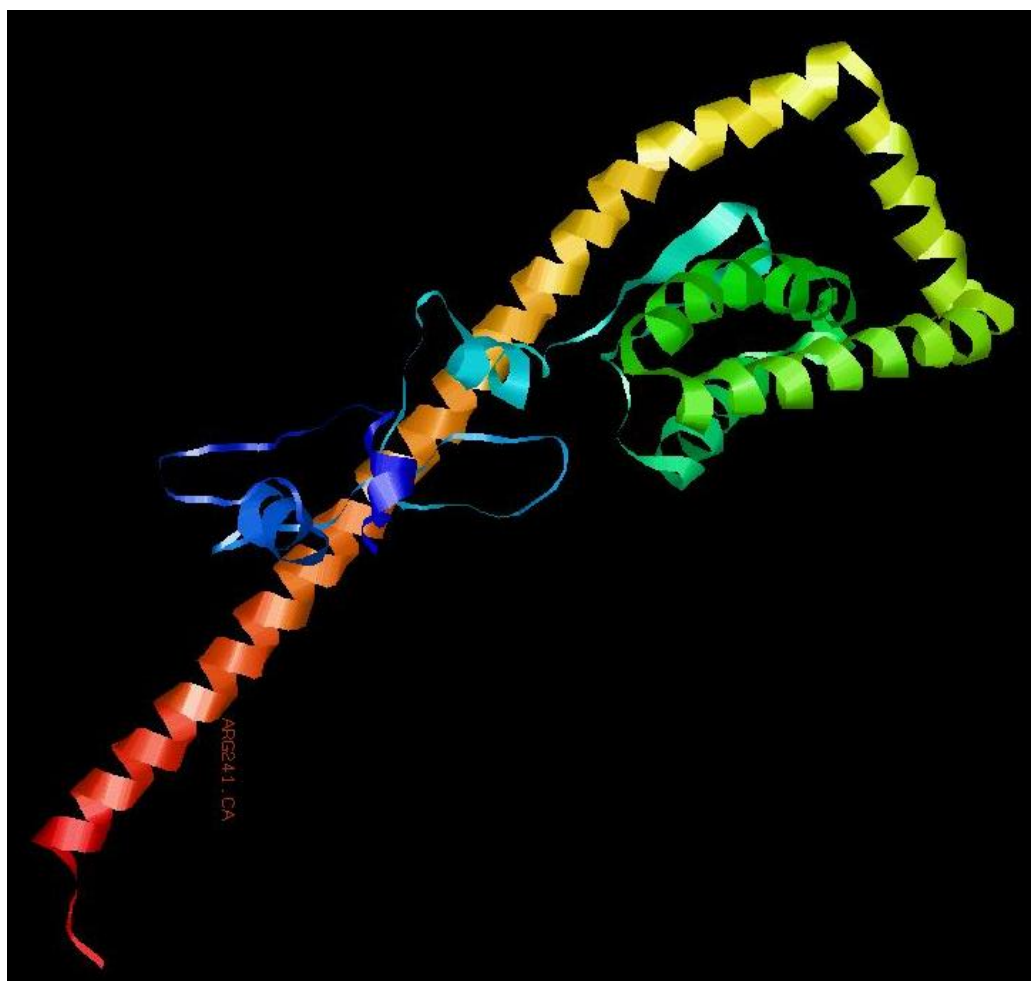
Model 4 – Wild Type



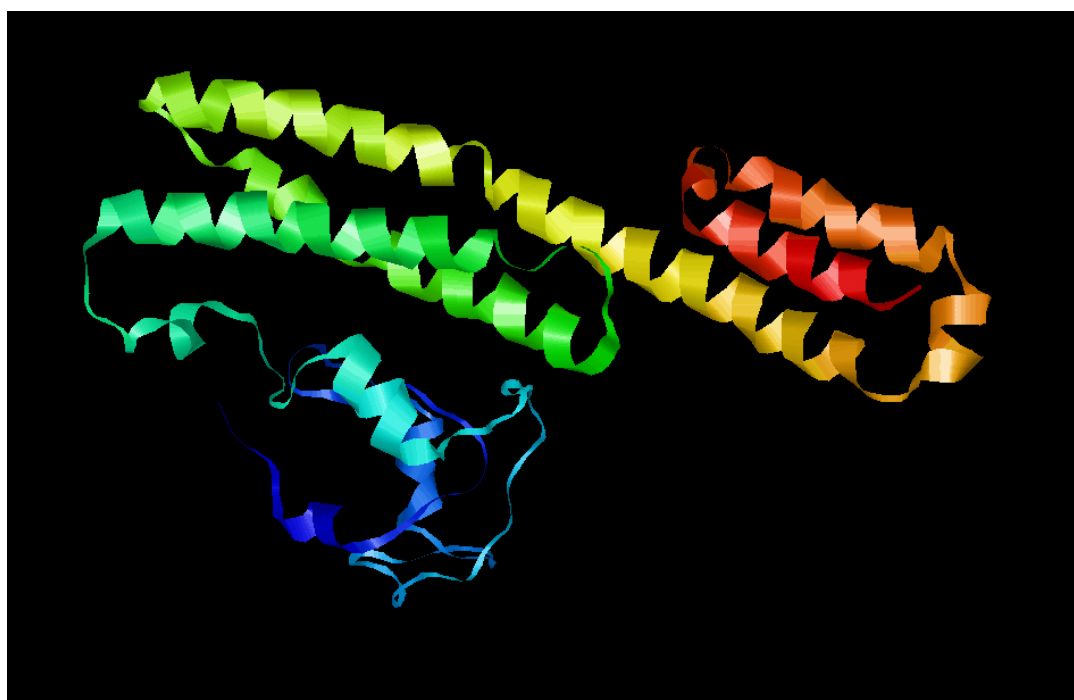
Model 4 – Mutant



Model 5 – Wild Type



Model 5 – Mutant



DNAAF3

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-267	Alignment	4ctmA_301	1-807	0.0431	
268-390	Alignment	3sumA_201	1-135	0.0604	Unknown function
391-506	Alignment	3opnA_301	1-208	0.3860	Structural genomics, unknown function
507-588	Alignment	3d87A_101	1-161	0.3765	Cytokine

Table 10.6 PDB Domains used for Ginzu Prediction when predicting DNAAF3p structure

Wild type sequence

MLPLLDSSKRAGTLGSGCGVPRVHSAALSREEGASRDIWRIKVMWARVMTTPAGSGSGFGSVSWWGLSPALDLQAE
 SPPVDPDSQADTVHSNPELDVLGSGVDGRHLLRTLRAKFWPRRRFNFFVLENNLEAVA R HMLIFSLALEEPEK
 MGLQERSETFLEVWGNALLRPPVAAFVRAQADLLAHLVPEPDRLEEQLPWLSLRALKFRERDALEAVFRFWAGGE
 KGPQAFPM SRLWDSRLRHYLGSRVDARRGVSDWDLRMKLHDRGAQVIHPQEFRRWRDTGVAFELRDSSAYHVPNR
 TLASGRLLSYRGERVAARGYWGDIATGPFVAFGIEADDESLLRTSNGQPVKTAGETITQHNVTLLRDVAWGRAR
 ATGGDLEEQQHAEGSPEPGTPAAPTPESTVHFLPLNSAQTLLHHKSCYNGRFQLLYVACGMVHLLIPELGACVAP
 GGNLIVELARYLVDVRQEQLQGFNTRVRELAQAAGFAPQTGARPSETFARFCKSQESALGNTVPAVEPGTPPLDI
 LAQPLEASNPALGLTQPLQGGTPHCEPCQLPSESPGSLSEVLAQPQGALAPPNCESDSKTGV

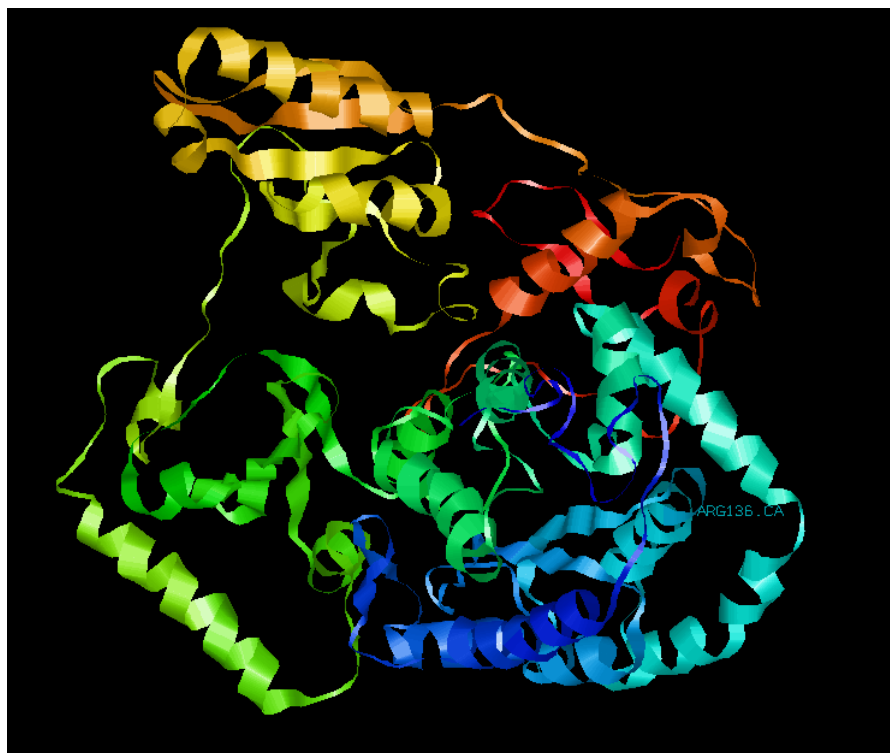
Mutant-type sequence (p.R136*)

MLPLLDSSKRAGTLGSGCGVPRVHSAALSREEGASRDIWRIKVMWARVMTTPAGSGSGFGSVSWWGLSPALDLQAE
 SPPVDPDSQADTVHSNPELDVLGSGVDGRHLLRTLRAKFWPRRRFNFFVLENNLEAVA

Other DNAAF3 Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

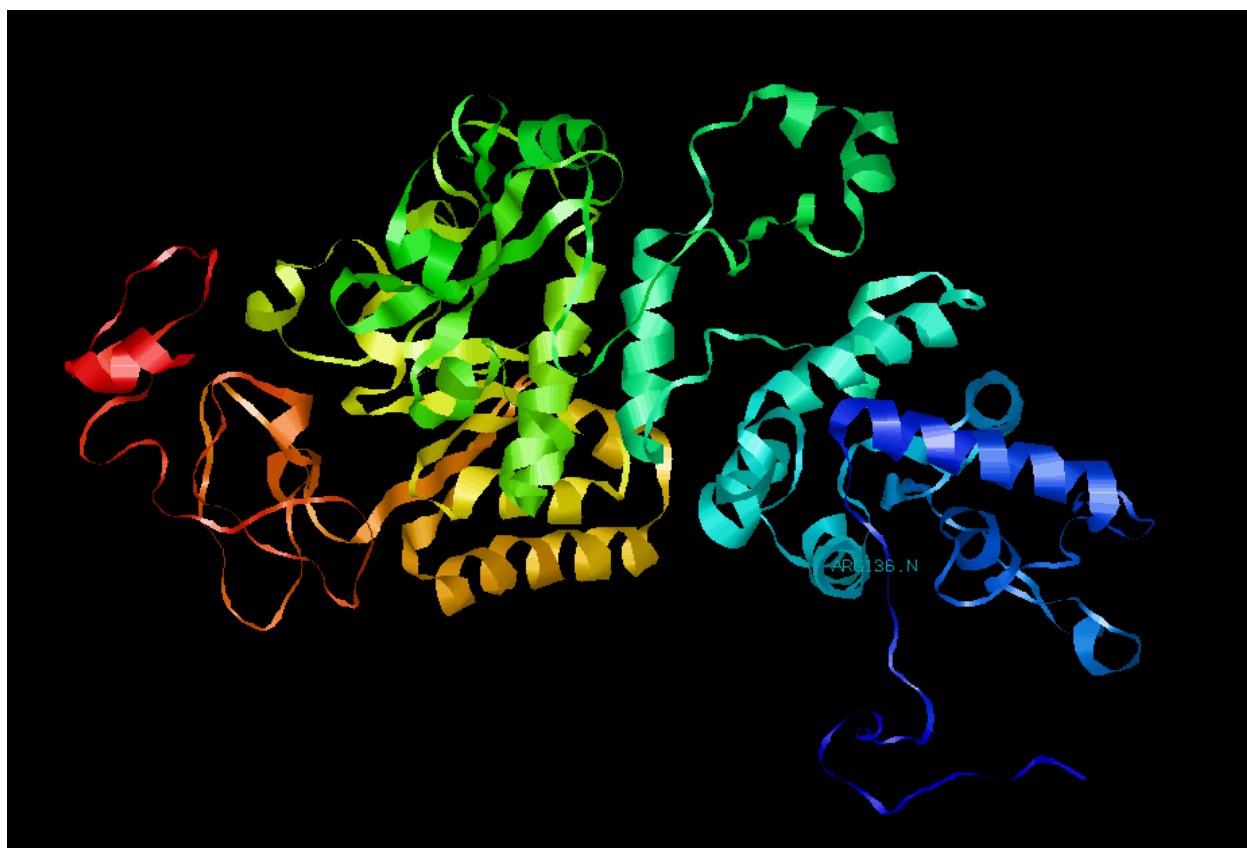
Model 2 – Wild Type



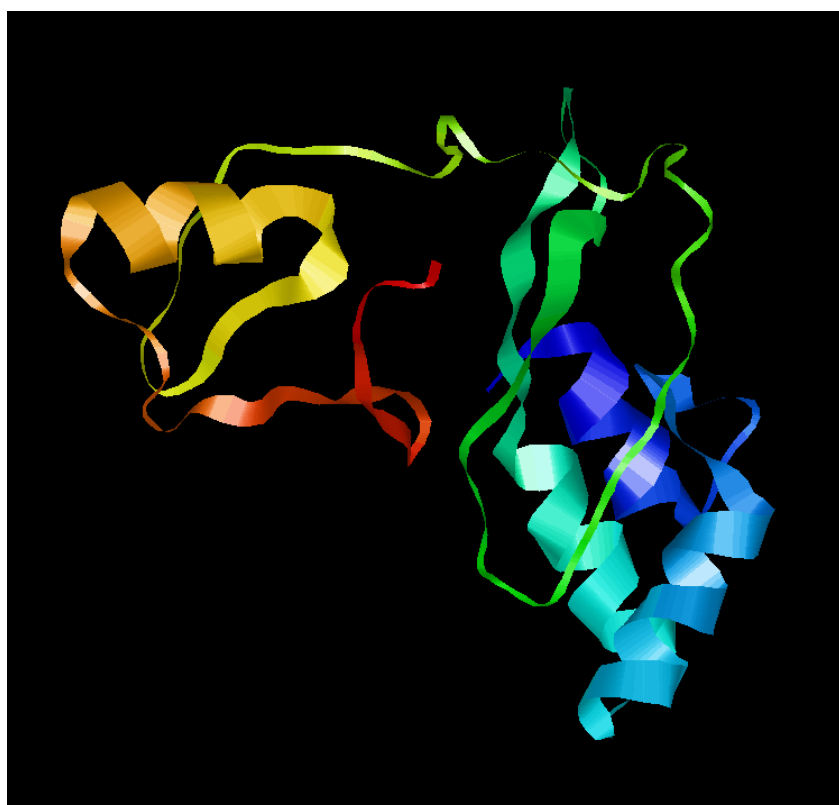
Model 2 – Mutant



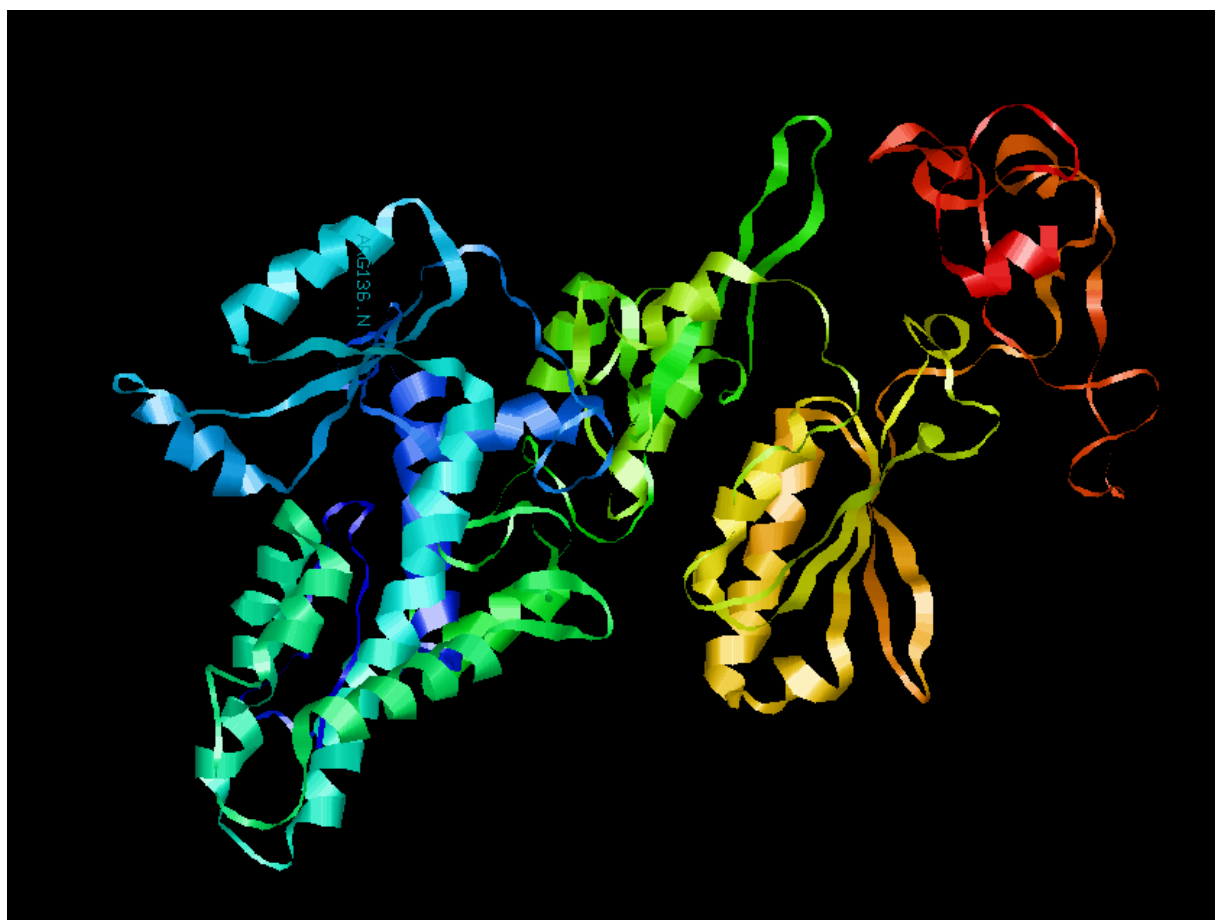
Model 3 – Wild Type



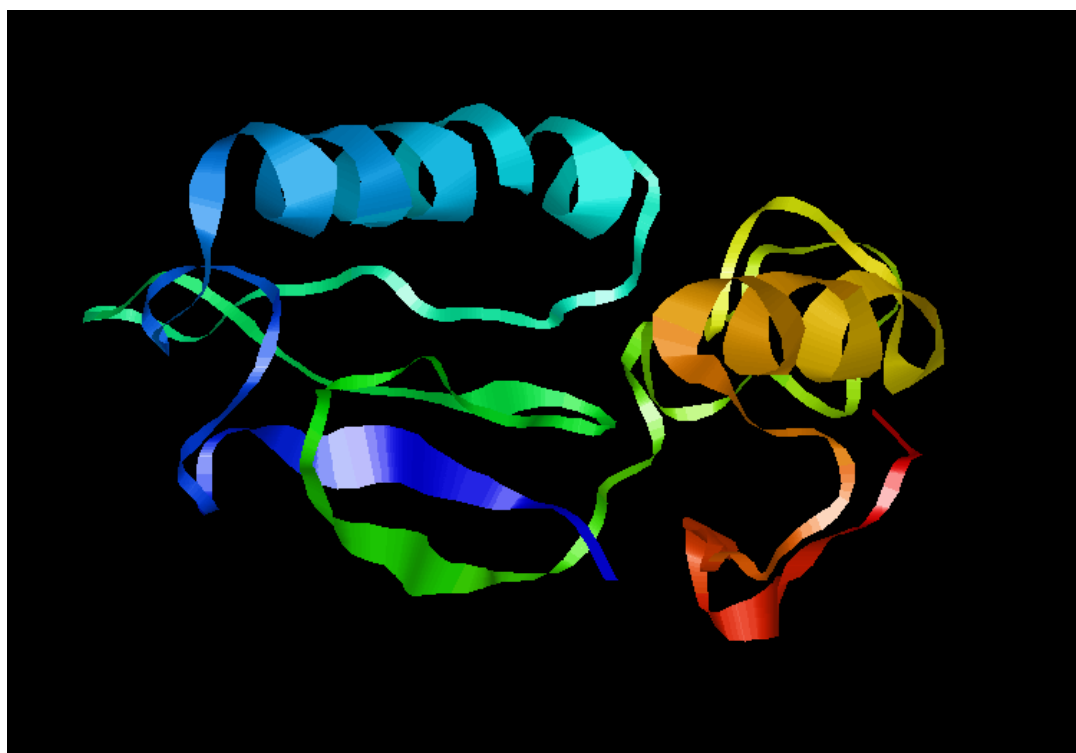
Model 3 – Mutant Type



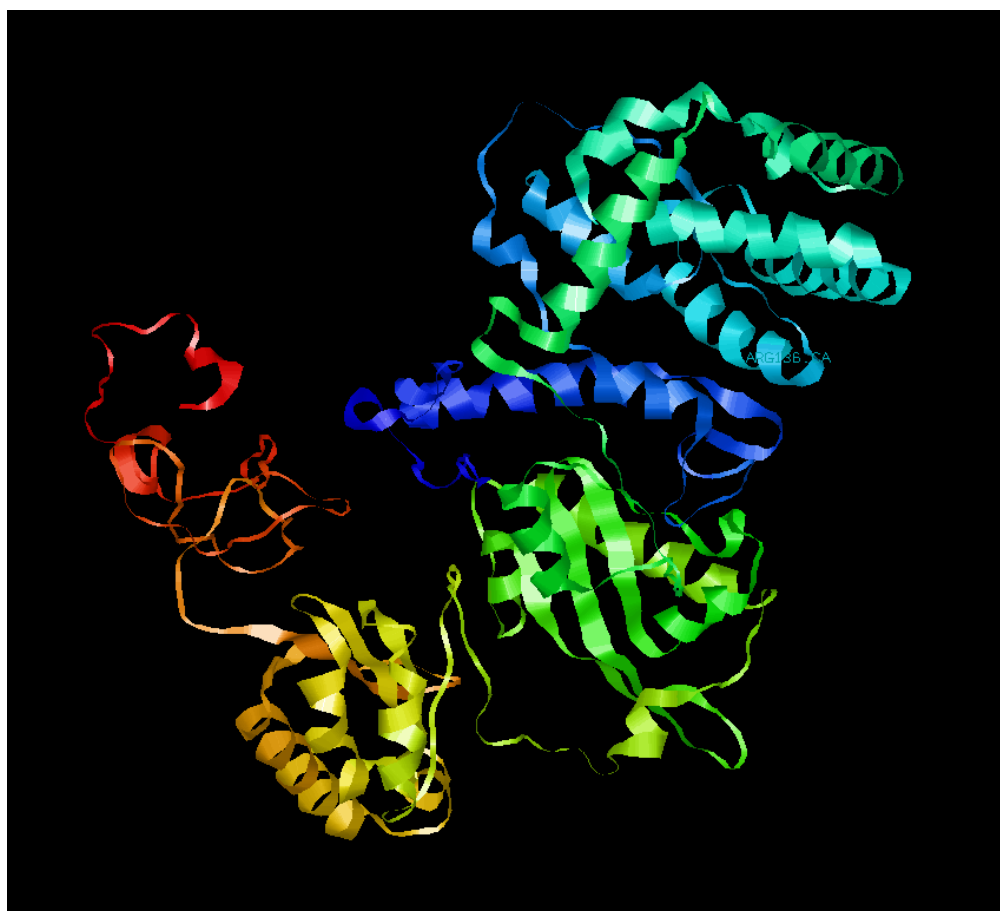
Model 4 - Wild Type



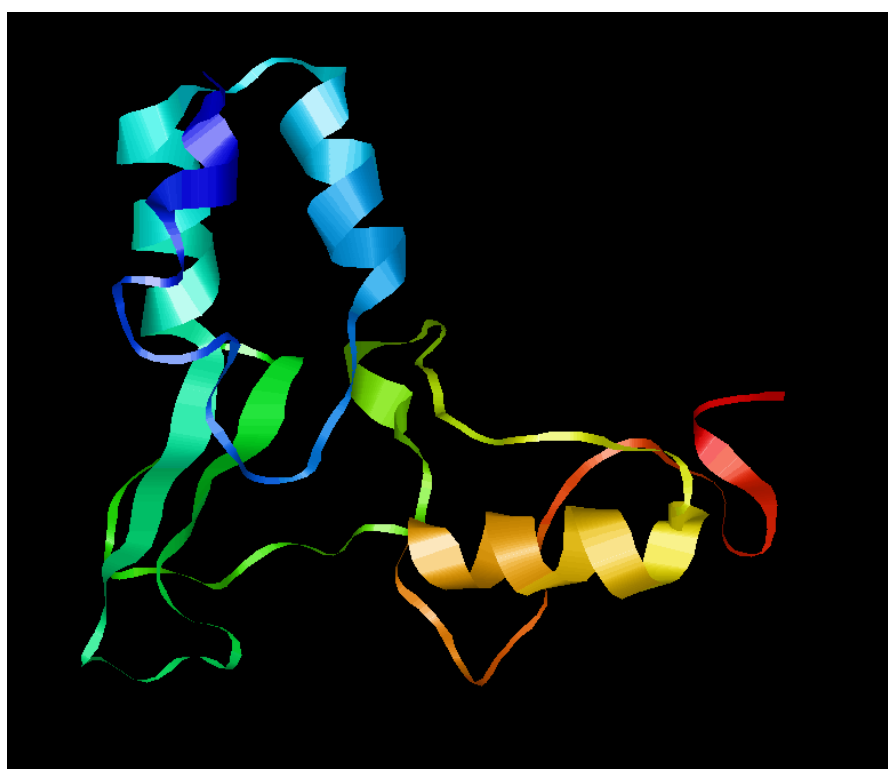
Model 4 - Mutant Type



Model 5 – Wild Type



Model 5 – Mutant Type



CCDC151

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-64	Alignment	3manA_101	1-297	0.2121	Hydrolase
65-329	Alignment	2w49A_301	1-277	0.3816	Contractile Protein
330-420	Alignment	1l8dA_301	1-103	0.1741	Replication
421-595	Alignment	2oevA_302	1-697	0.0838	Protein Transport

Table 10.7 PDB Domains used for Ginzu Prediction when predicting CCDC151p structure

Wild type sequence

MTSPLCRAASANALPPQDQASTPSSRVKGREASGKPSHLRGKGTAQAWTPGRSKGGSFHRGAGKPSVHSQVAELH
KKIQLLEGDRKAFFESSQWNIKKNQETISQLRKETKALELKLLDLLKGDEKVVQAVIREWKWEKPYLKNRTGQAL
EHLDHRLREKVKQONALRHQVVLRRRLEELQLQHSLRLLEMAEAQNRHTEVAKTMRNLENRLEKAQMKAQEAH
ITSVYLQLKAYLMDESLNLENRLDSMEAEEVVRTKHELEALHVVNQEALNARDIAKNQLQYLEETLVRERKKRERY
ISECKKRAEEKLENERMERKTHREHLLQSDDTIQDSLHAKKEELRQRWSMYQMEVIFGKVKDATGTDETHSLV
RRFLAQGDTFQLETLKSENEQTLVRLKQEKQQLQRELEDLKYSGEATLVSQQKLQAEAQERLKKKEERRHAEAKD
QLERALRAMQVAKDSLEHLASKLIHITVEDGRFAGKELDPQADNYVPNLLGLVEEKLLKLQAQLQGHVQEMLCH
IANREFLASLEGRLPEYNTRIALPLATSKDKFFDEESEEEDNEVVTRASLKIRSQKLIESHKKHRRSRRS

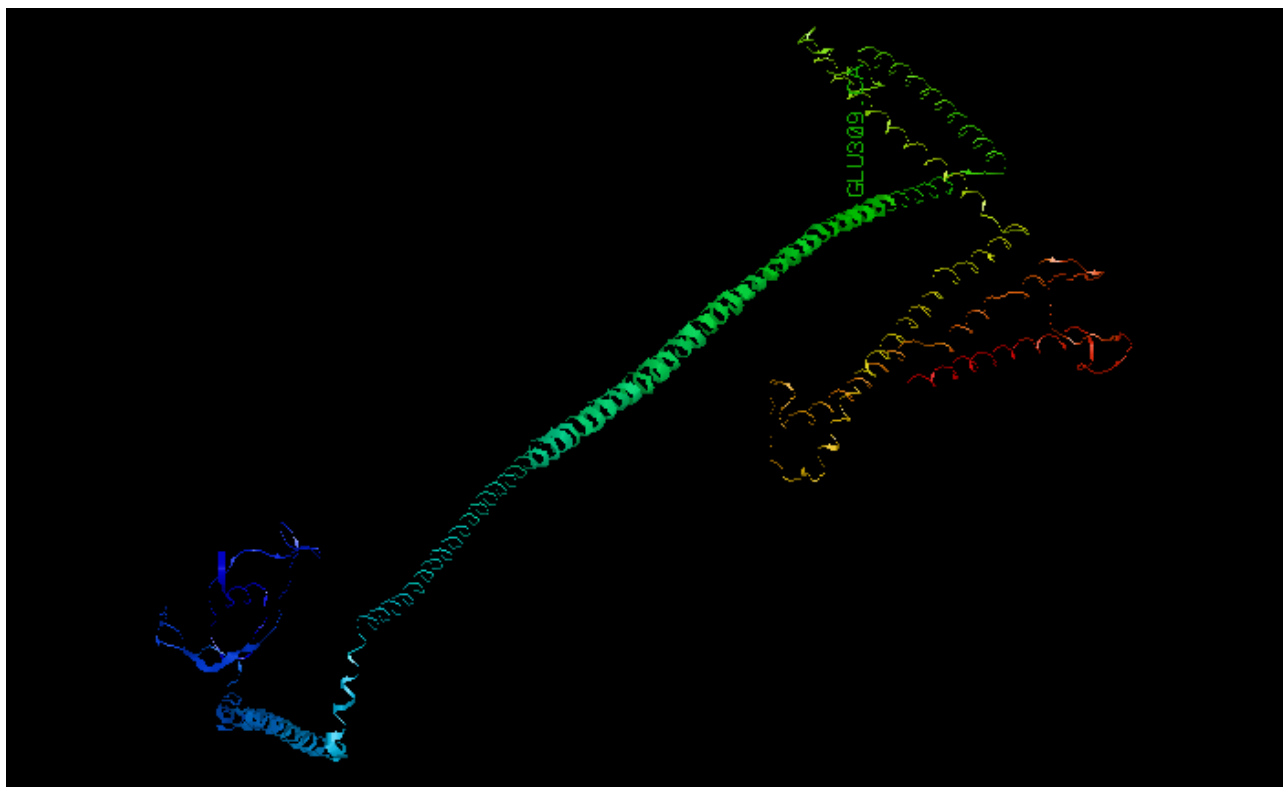
Mutant-type sequence (p.E309*)

MTSPLCRAASANALPPQDQASTPSSRVKGREASGKPSHLRGKGTAQAWTPGRSKGGSFHRGAGKPSVHSQVAELH
KKIQLLEGDRKAFFESSQWNIKKNQETISQLRKETKALELKLLDLLKGDEKVVQAVIREWKWEKPYLKNRTGQAL
EHLDHRLREKVKQONALRHQVVLRRRLEELQLQHSLRLLEMAEAQNRHTEVAKTMRNLENRLEKAQMKAQEAH
ITSVYLQLKAYLMDESLNLENRLDSMEAEEVVRTKHELEALHVVNQEALNARDIAKNQLQYLEETLVRERKKRERY
ISECKKRA

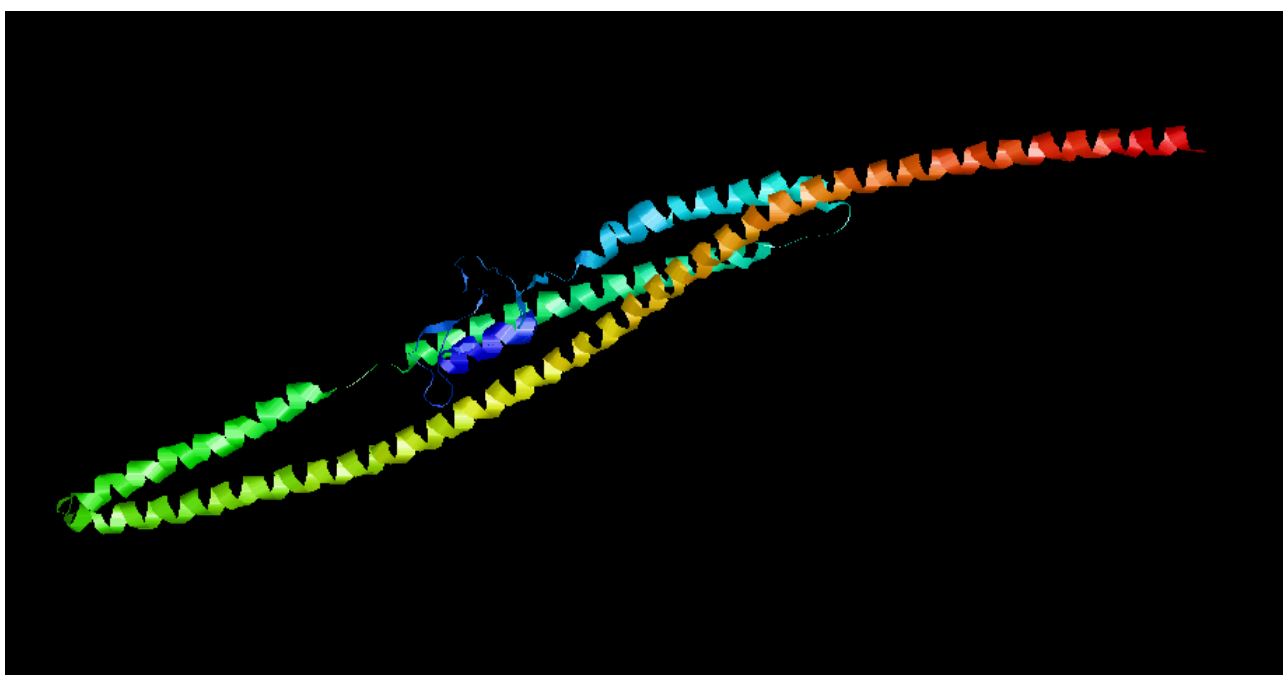
Other CCDC151 Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

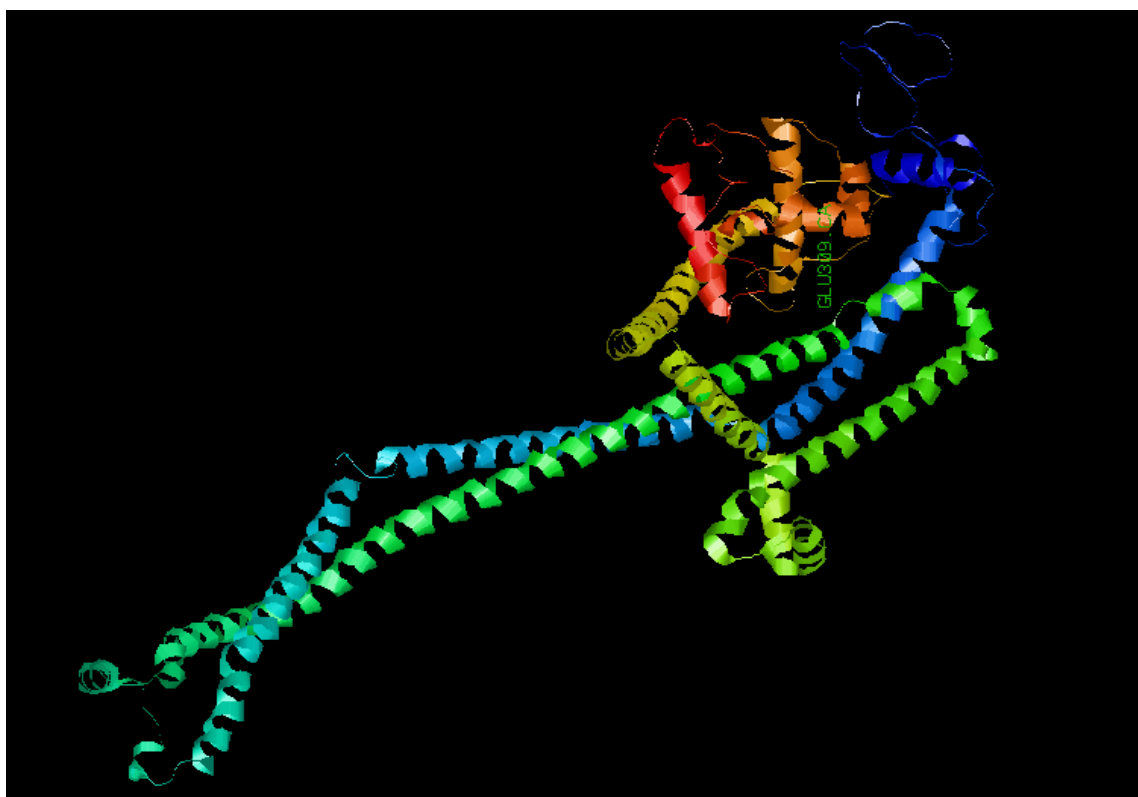
Model 2 – Wild Type



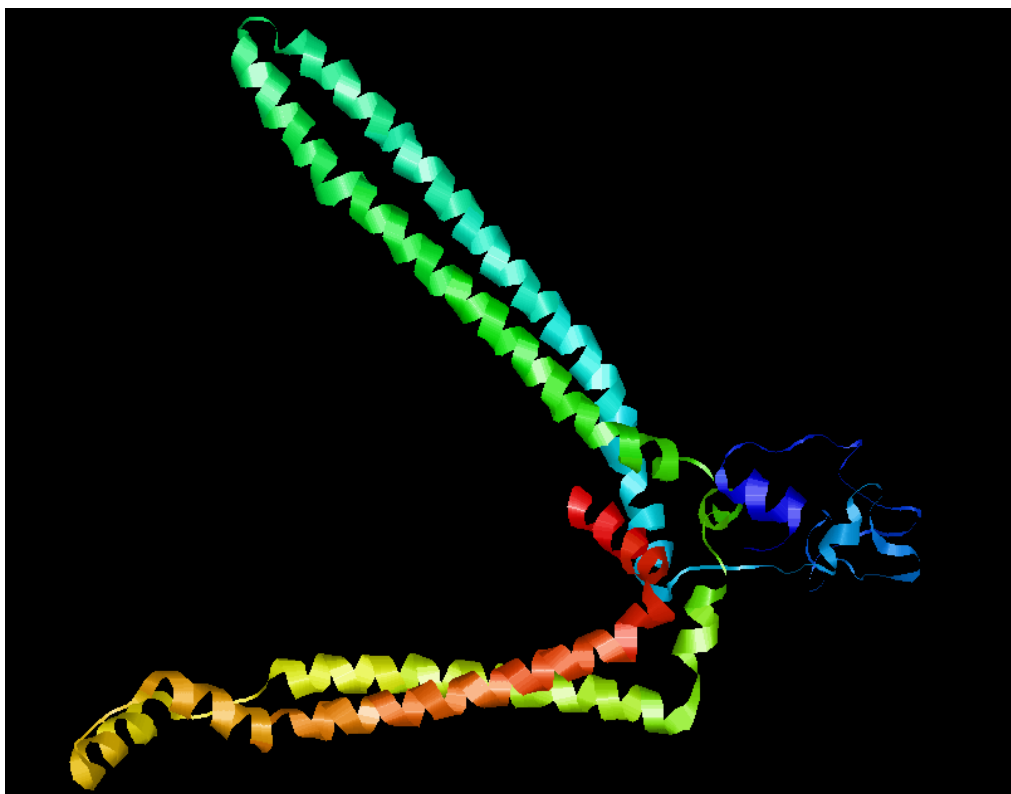
Model 2 – Mutant Type



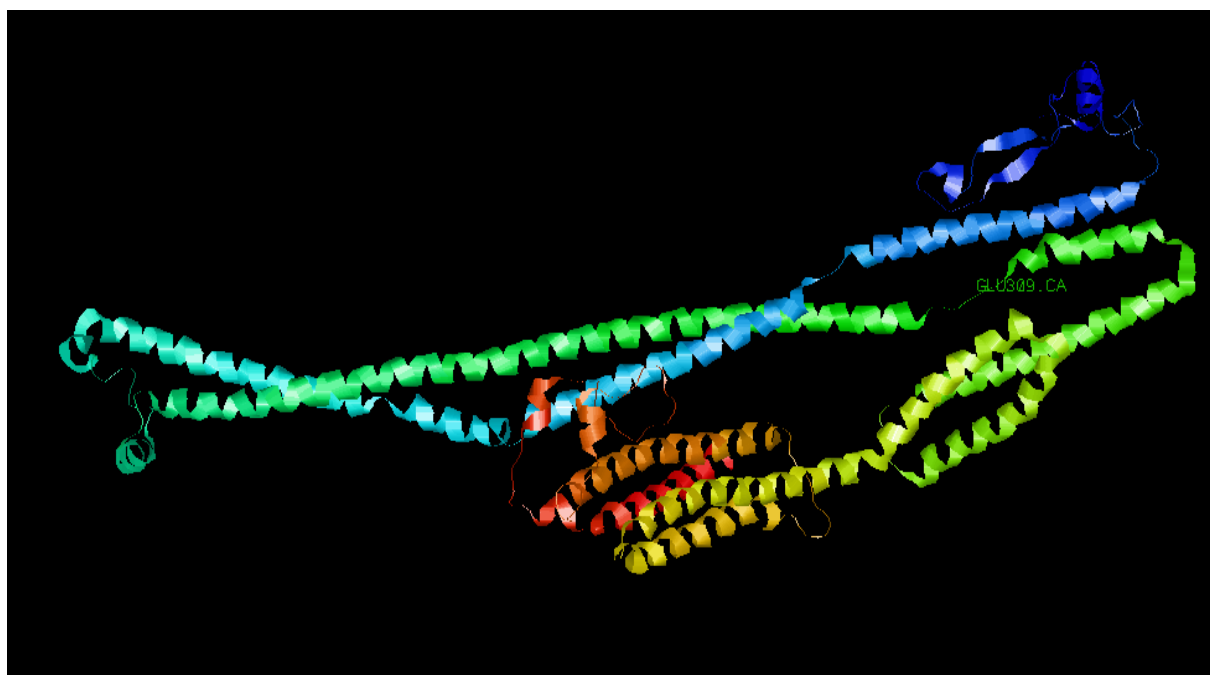
Model 3 – Wild Type



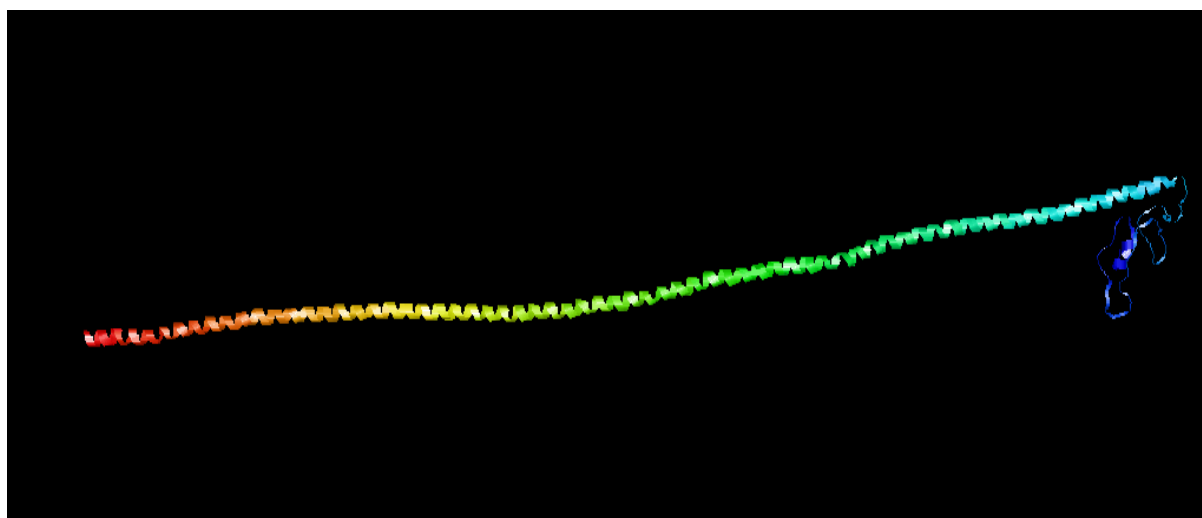
Model 3 – Mutant Type



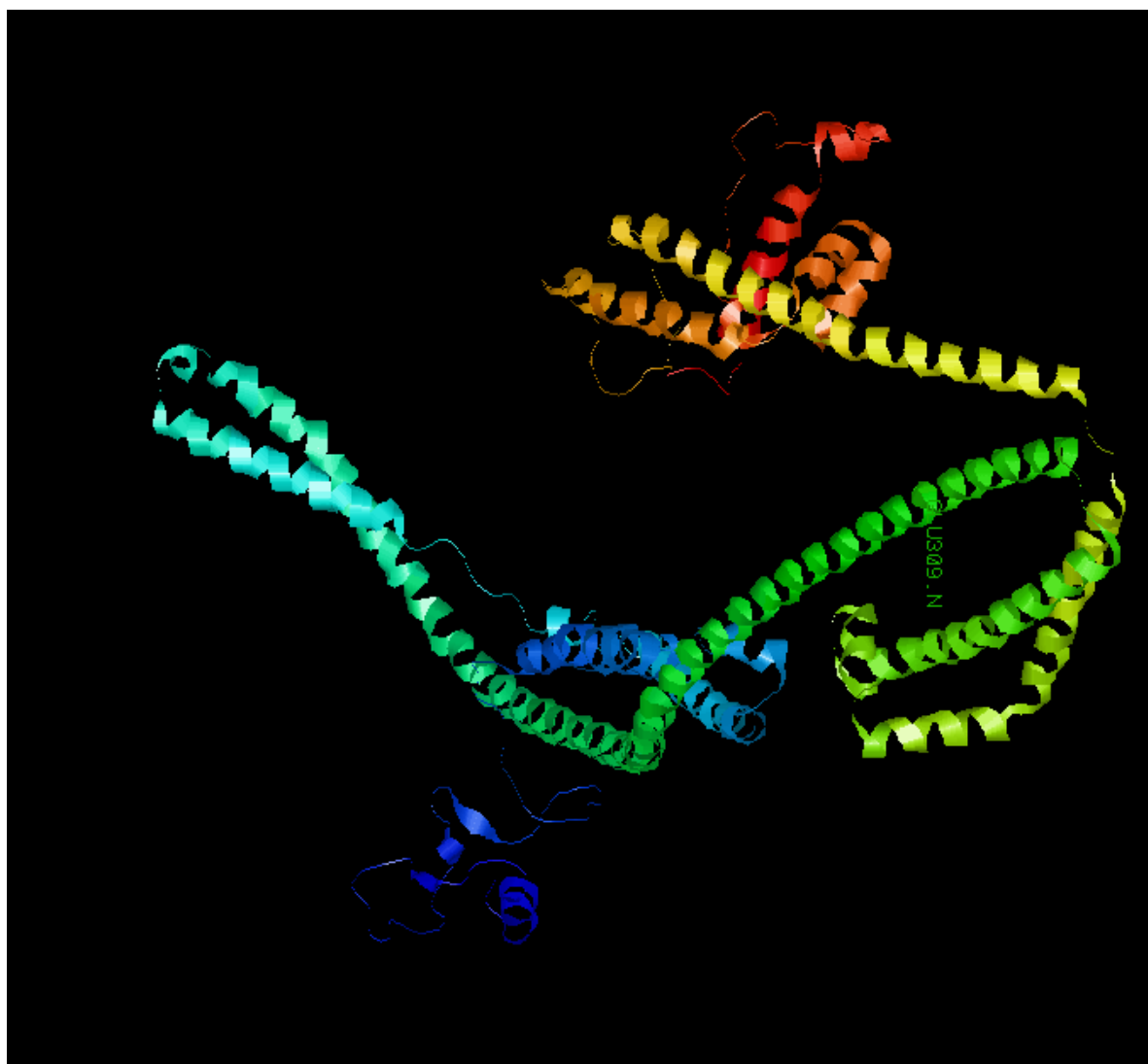
Model 4 – Wild Type



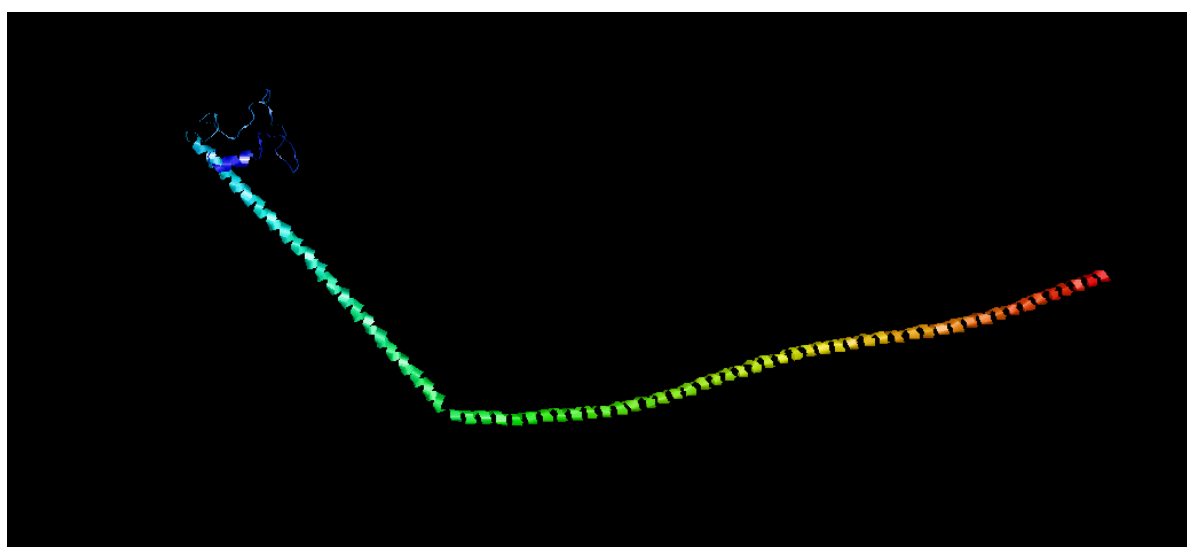
Model 4 – Mutant Type



Model 5 – Wild Type



Model 5 – Mutant Type



CTSC

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-130	Alignment	1jqpA_201	1-348	0.8846	Hydrolase
131-439	Alignment	3pdfA_201	1-351	0.5725	Hydrolase/hydrolase inhibitor

Table 10.8 PDB Domains used for Ginzu Prediction when predicting CTSCp structure

Wild type sequence

MGAGPSLLLAALLLLLSGDGAVRCDTPANCTYLDLLGTWVFQVGSSGSQRDVNCSVMGPQEKKVVVYLQKLD TAY
 DDLGNSGHFTIIYNQGFEIVLNDYKWF AFFKYKEEGSKVTTCNETMTGWVHDVLGRNWACFTGKKVGTASENVY
 VNIAHLKNSQEKYSNRLYKYDHNFKAINAIQKSWTATTYMEYETLT LGDMIRRS GGHSRKIPRKPAPLTAEIQ
 QKILHLPTS WDWRNVHGINFVSPVRNQASC GSCY SFASMGMLEARIRILTNN SQTPILSPQEVVSCSQYAQGCE **C**
 GFPLYIAGKYAQDFGLVEEACFPYTGT DSPCKMKEDCFRYY SSEYHYVGGFYGGCNEALMKLELVHHGPM AVAFE
 VYDDFLHYKKG IYHHTGLRDPFNP FELTNH AVLLVGYGTDSASGMDYWIVKNSWGTGWGENGYFRIRRG TDECAI
 ESIAVAATPIPKL

Mutant-type sequence (p.G300D)

MGAGPSLLLAALLLLLSGDGAVRCDTPANCTYLDLLGTWVFQVGSSGSQRDVNCSVMGPQEKKVVVYLQKLD TAY
 DDLGNSGHFTIIYNQGFEIVLNDYKWF AFFKYKEEGSKVTTCNETMTGWVHDVLGRNWACFTGKKVGTASENVY
 VNIAHLKNSQEKYSNRLYKYDHNFKAINAIQKSWTATTYMEYETLT LGDMIRRS GGHSRKIPRKPAPLTAEIQ
 QKILHLPTS WDWRNVHGINFVSPVRNQASC GSCY SFASMGMLEARIRILTNN SQTPILSPQEVVSCSQYAQGCE **D**
 GFPLYIAGKYAQDFGLVEEACFPYTGT DSPCKMKEDCFRYY SSEYHYVGGFYGGCNEALMKLELVHHGPM AVAFE
 VYDDFLHYKKG IYHHTGLRDPFNP FELTNH AVLLVGYGTDSASGMDYWIVKNSWGTGWGENGYFRIRRG TDECAI
 ESIAVAATPIPKL

Wild type CTSC PDB (at position 300) – Model 1

ATOM	4659	N	GLY	A	300	-30.119	30.410	4.127	1.00	3.37
ATOM	4660	CA	GLY	A	300	-29.096	30.609	3.102	1.00	2.99
ATOM	4661	C	GLY	A	300	-28.252	29.376	2.800	1.00	3.54
ATOM	4662	O	GLY	A	300	-28.378	28.347	3.457	1.00	7.98
ATOM	4663	H	GLY	A	300	-30.841	29.721	3.974	1.00	4.04
ATOM	4664	1HA	GLY	A	300	-29.578	30.907	2.171	1.00	3.59
ATOM	4665	2HA	GLY	A	300	-28.416	31.396	3.426	1.00	3.59

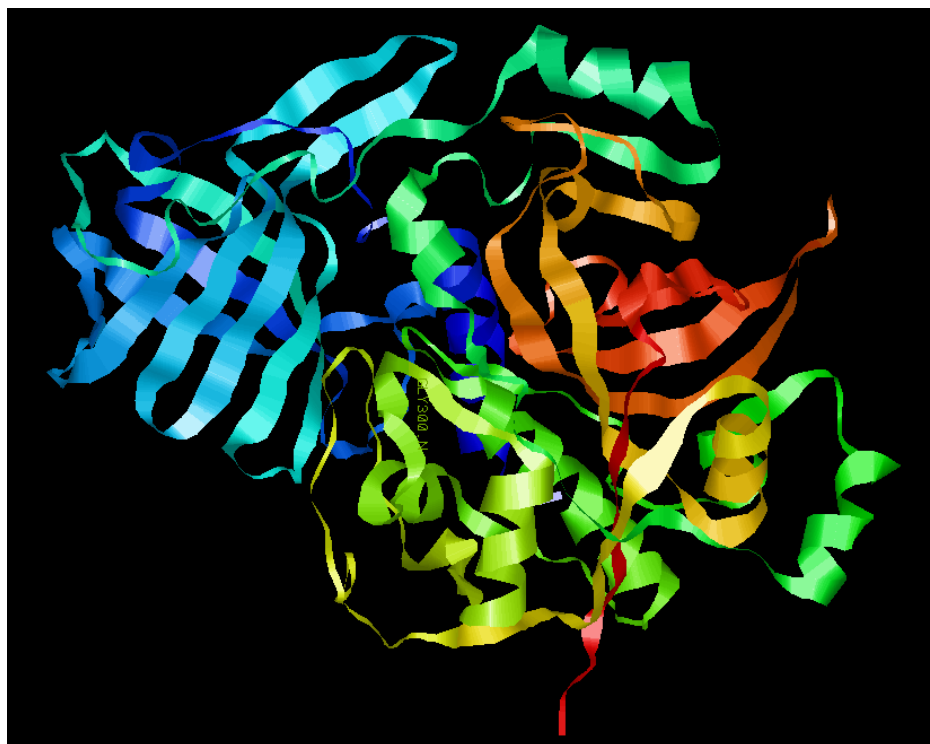
Mutant CTSC PDB (at position 300) – Model 1

ATOM	4659	N	ASP	A	300	-13.507	-7.245	-9.960	1.00	1.62
ATOM	4660	CA	ASP	A	300	-13.676	-5.810	-10.229	1.00	0.55
ATOM	4661	C	ASP	A	300	-12.878	-5.256	-11.419	1.00	0.81
ATOM	4662	O	ASP	A	300	-12.103	-5.959	-12.067	1.00	3.15
ATOM	4663	CB	ASP	A	300	-13.300	-4.985	-8.996	1.00	0.83
ATOM	4664	CG	ASP	A	300	-14.285	-5.096	-7.840	1.00	0.83
ATOM	4665	OD1	ASP	A	300	-15.363	-5.600	-8.050	1.00	0.83
ATOM	4666	OD2	ASP	A	300	-13.900	-4.822	-6.729	1.00	0.83
ATOM	4667	H	ASP	A	300	-14.246	-7.884	-10.217	1.00	1.94
ATOM	4668	HA	ASP	A	300	-14.713	-5.604	-10.496	1.00	0.66
ATOM	4669	1HB	ASP	A	300	-12.291	-5.182	-8.632	1.00	0.99
ATOM	4670	2HB	ASP	A	300	-13.346	-3.980	-9.416	1.00	0.99

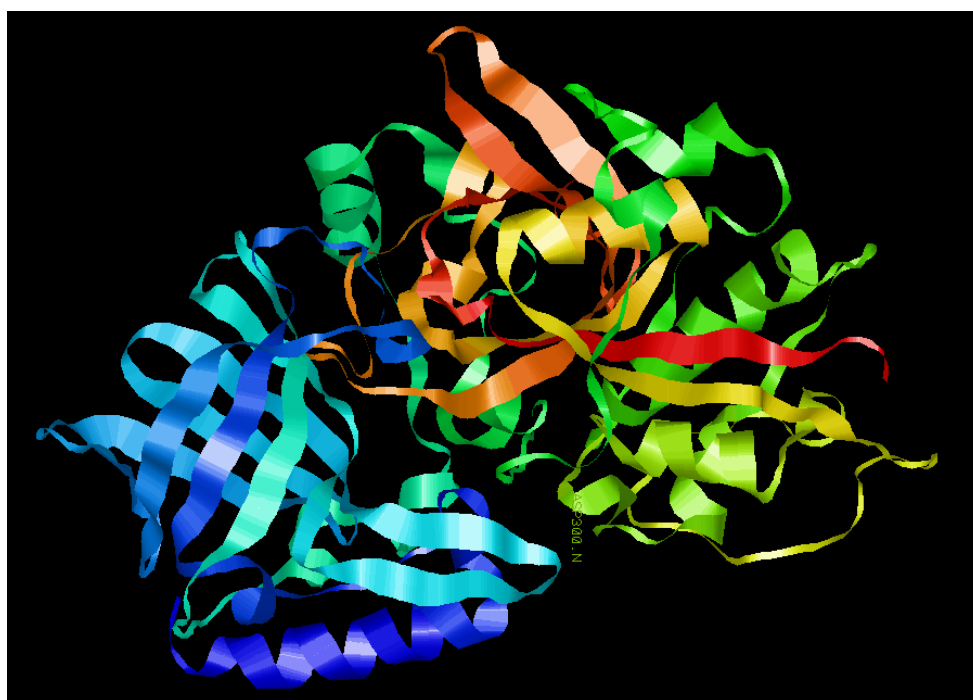
Other CTSC Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

Model 2 – Wild Type



Model 2 – Mutant



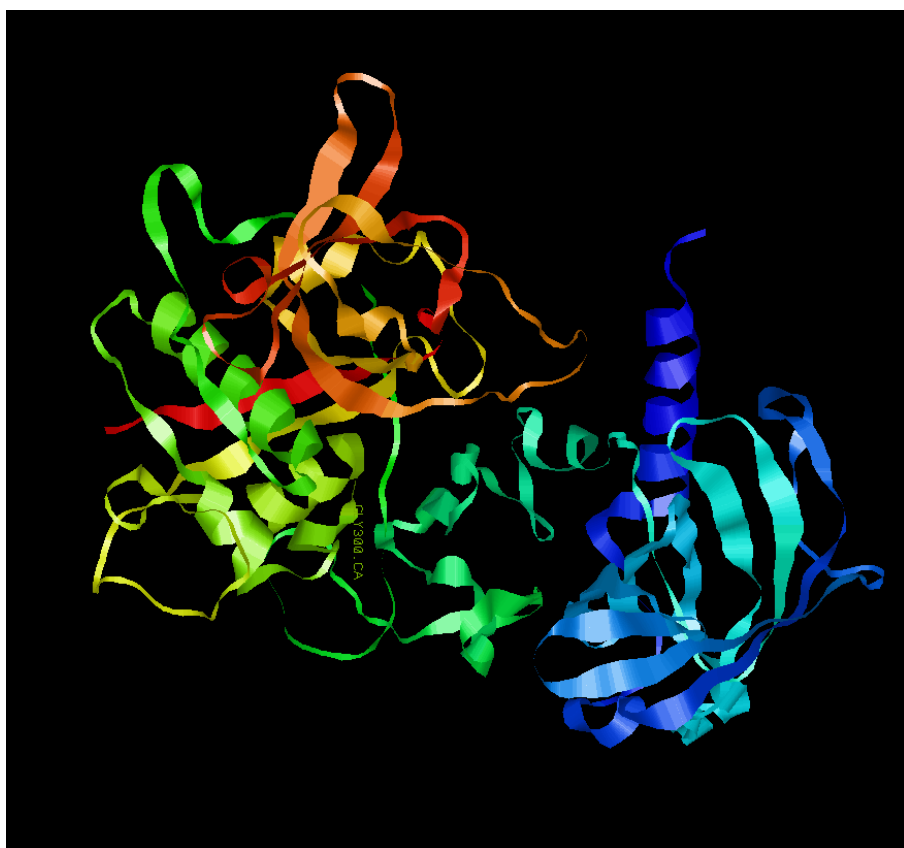
Model 3 – Wild Type



Model 3 – Mutant



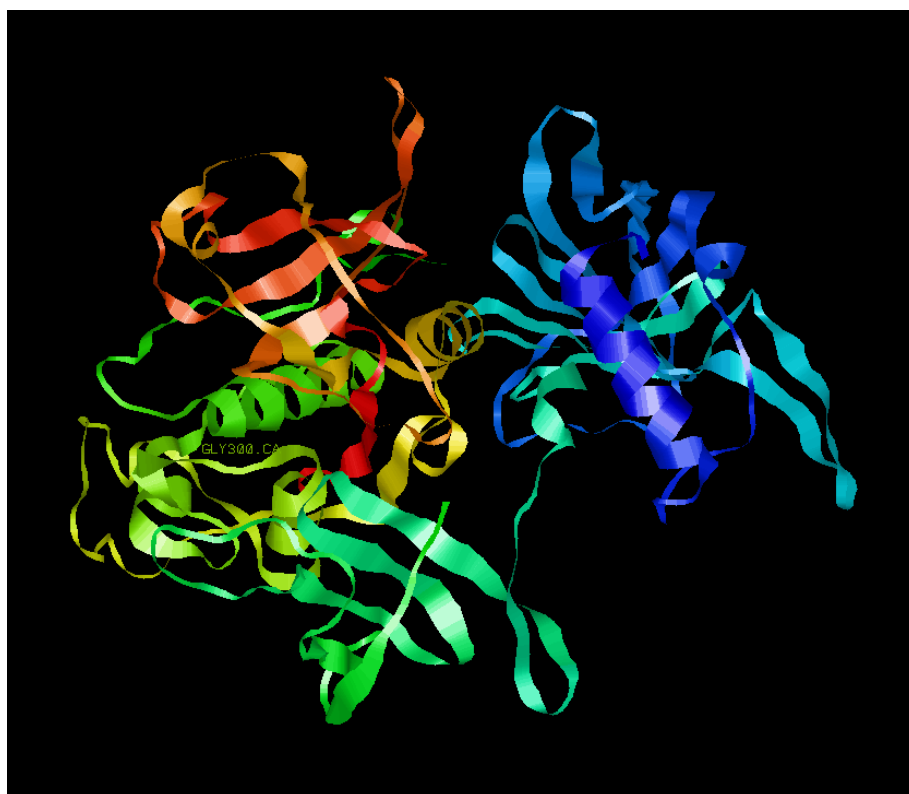
Model 4 - Wild Type



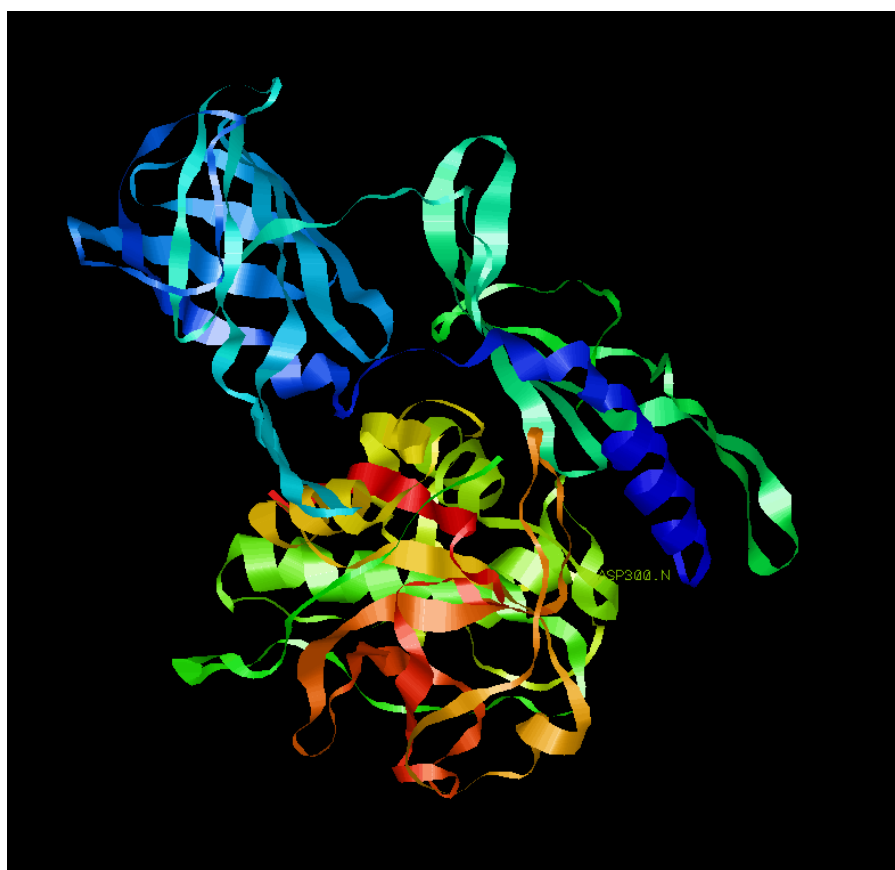
Model 4 - Mutant



Model 5 – Wild Type



Model 5 – Mutant



9.6. Appendices for Chapter 5 - Intellectual disability analyses

9.6.1. Protein structure modelling

ADAT3

Domains used for Ginzu Prediction

Domain Span	Source	Reference Parent	Parent Span	Confidence	Annotations
1-160	alignment	4lmzA_301	1-176	0.044800	
161-351	alignment	3dh1A_202	1-162	0.723400	Hydrolase

Table 10.9 PDB Domains used for Ginzu Prediction when predicting ADAT3p structure

Wild type sequence

MEPAPGLVEQPKCLEAGSPEPEPAPWQALPVLSEKQSGDVELVLAYAAPVLDKRQTSRLLKEVSALHPLPAQPHL
KVRPSPRDAGSPHALEMLLCLAGPASGPRSLAELLPRPAVDPRGLGQPFLVPYPARPPLTRGQFEEARAHWPTSF
HEDKQVTSALAGRLFSTQERAAMQSHMERAVWAARRAAARGLRAVGAVVVDPASDRVLATGHDCSCADNPLLHAV
MVCVDLVARGQGGRGTYDFRPFACSFAPAAAPQAVRAGAVRKLDADDEDGLPYLCTGYDLYVTREPCAMCAMLVH
ARILRVFYGAPSPDGALGTRFRIHARPDNLNHRFQVFRGVLEEQCRWLDPDT

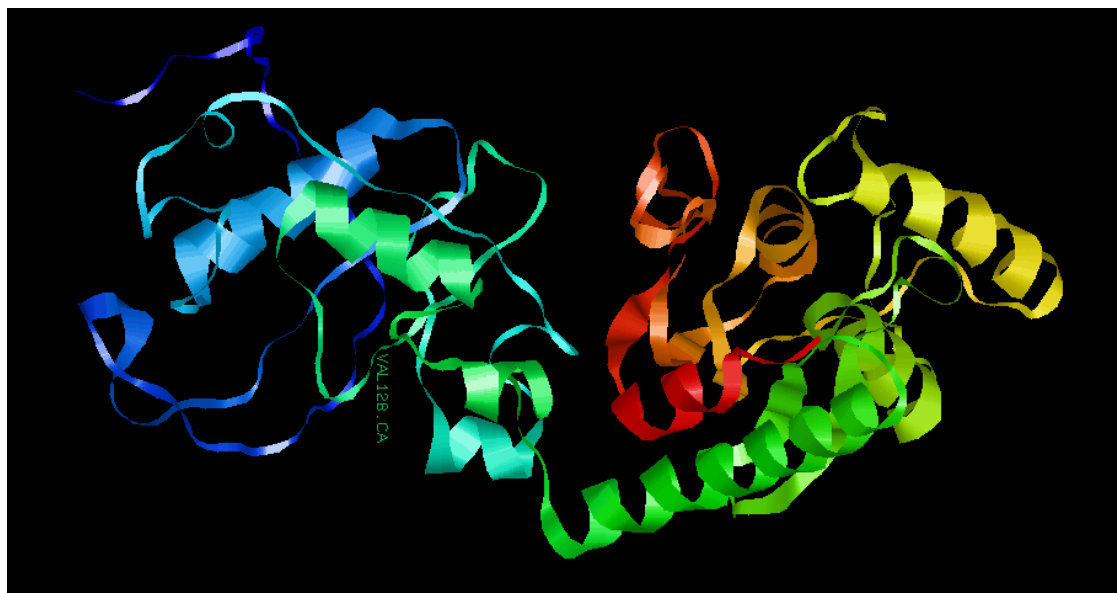
Mutant-type sequence (p.V128M)

MEPAPGLVEQPKCLEAGSPEPEPAPWQALPVLSEKQSGDVELVLAYAAPVLDKRQTSRLLKEVSALHPLPAQPHL
KVRPSPRDAGSPHALEMLLCLAGPASGPRSLAELLPRPAVDPRGLGQPFLVPMYPARPPLTRGQFEEARAHWPTSF
HEDKQVTSALAGRLFSTQERAAMQSHMERAVWAARRAAARGLRAVGAVVVDPASDRVLATGHDCSCADNPLLHAV
MVCVDLVARGQGGRGTYDFRPFACSFAPAAAPQAVRAGAVRKLDADDEDGLPYLCTGYDLYVTREPCAMCAMLVH
ARILRVFYGAPSPDGALGTRFRIHARPDNLNHRFQVFRGVLEEQCRWLDPDT

Other ADAT3 Protein Structure predictions

The following figures present the top 5 most likely structures (except the first model which was presented in respective section) of the wild type and mutant type proteins.

Model 2 – Wild Type



Model 2 – Mutant



Model 3 – Wild Type



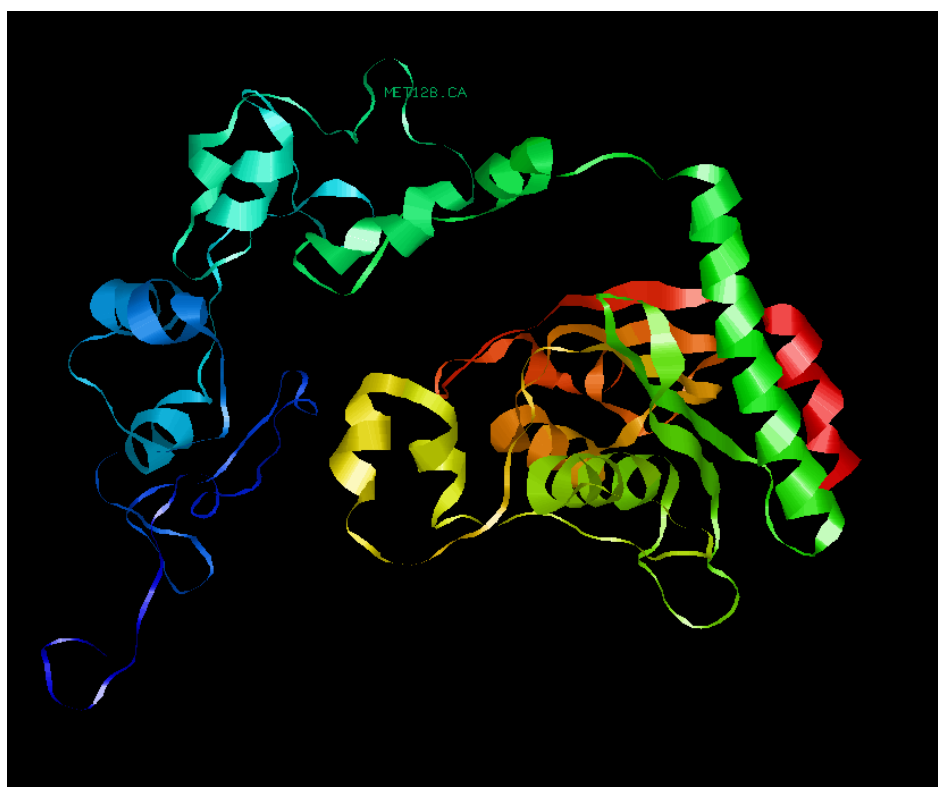
Model 3 – Mutant



Model 4 – Wild Type



Model 4 – Mutant



Model 5 – Wild Type



Model 5 – Mutant



9.6.2. Autozygosity mapping within family

Individual ID	26	27	29	30	31	33
LROHs	Chr1 6555538 7253934	Chr1 18082615 33792054	Chr1 6555538 7253934	Chr1 102518054 103904580	Chr1 11694927 17383290	Chr1 72017 2396747
	Chr1 102518054 103904580	Chr1 44738310 46267611	Chr1 41925922 42654883	Chr1 236018105 241138539	Chr1 30761117 36630732	Chr1 12427744 14009289
	Chr1 168019324 168639127	Chr1 207904306 208553518	Chr1 44738310 46267611	Chr1 241152818 246599908	Chr1 87355953 94975572	Chr1 19075504 22976971
	Chr1 207904306 208553518	Chr1 236018105 246599908	Chr1 70761270 71680128	Chr2 129631363 130740891	Chr1 97778810 98147098	Chr1 49004662 50431666
	Chr1 236018105 246599908	Chr2 1799067 2857590	Chr1 75857708 76720632	Chr2 135055110 136693406	Chr1 113038204 164171324	Chr1 87355953 94975572
	Chr2 47424810 47636059	Chr2 38479811 39195622	Chr1 205371523 208856837	Chr2 210013973 211250020	Chr1 192517667 193452046	Chr1 97778810 98147098
	Chr2 100353819 100918527	Chr2 47443538 47563733	Chr1 236018105 246599908	Chr2 233827154 234272602	Chr1 197955159 198738087	Chr1 113038204 164171324
	Chr2 210013973 211250020	Chr2 62345079 66853170	Chr2 1799067 2857590	Chr3 36495 1319139	Chr2 57280437 58083737	Chr1 196362503 214739428
	Chr2 233827154 234272602	Chr2 96287100 97989129	Chr2 64224090 66853170	Chr3 37005233 37068068	Chr2 107976696 108992589	Chr1 239641490 240264827
	Chr3 37005233 37068068	Chr2 129631363 130740891	Chr2 70537959 74031588	Chr3 45278317 49936784	Chr2 123218967 123791837	Chr2 47443538 47563733
	Chr3 64944490 96576822	Chr2 135055110 136693406	Chr2 96287100 97989129	Chr3 109288024 110445946	Chr2 134360566 137718603	Chr2 54409920 54925365
	Chr3 122379374 123192776	Chr2 233827154 234272602	Chr2 133109103 134336738	Chr3 122010027 123371778	Chr2 150176741 163333833	Chr2 151975605 153007734
	Chr4 25592985 27515927	Chr3 36495 1166638	Chr2 140197294 140919097	Chr3 141656975 142478588	Chr2 166357038 166814018	Chr2 213288968 215605474
	Chr4 38217325 60055013	Chr3 1317013 5472517	Chr2 147518397 148694053	Chr4 11521723 12223072	Chr3 5425470 6111434	Chr2 233783621 237860559
	Chr4 61648868 75967145	Chr3 141656975 142478588	Chr2 227769642 234220548	Chr4 13286528 13897404	Chr3 38661196 46634864	Chr3 3701784 5010285
	Chr4 147095001 149148251	Chr4 4077676 11021185	Chr2 234247389 240619095	Chr4 55925876 56685790	Chr3 50163890 51986002	Chr3 36940162 37123962
	Chr4 154110096 164019208	Chr4 25592985 27515927	Chr3 9876479 16188684	Chr4 99746639 110201134	Chr3 86504948 87579189	Chr3 44116161 45137320
	Chr4 180251050 180849310	Chr4 55925876 56685790	Chr3 62858090 77206651	Chr5 44667765 49897756	Chr3 120584743 122015901	Chr3 86504948 87579189
	Chr4 182970120 187925108	Chr4 140053341 151143887	Chr3 109288024 110445946	Chr5 71275894 73555480	Chr3 195303912 197044336	Chr3 120584743 122015901

	Chr5	30391830	Chr5	80564	Chr3	122010973	Chr5	111463668	Chr4	61566	Chr3	122379374
		73616923		3037374		123371778		112004635		3968061		123192776
	Chr5	108069506	Chr5	63153544	Chr4	10808450	Chr5	137419023	Chr4	4077676	Chr3	156394896
		115291752		73616923		25615024		139488889		5130953		157219098
	Chr5	137872173	Chr5	79273243	Chr4	55925876	Chr5	168136033	Chr4	6568947	Chr3	169282596
		139153625		79878733		56685790		169373200		7794035		172584250
	Chr5	161628109	Chr5	95845102	Chr4	117858250	Chr5	171511489	Chr4	11153240	Chr4	70553301
		168041434		97191558		119381243		172109066		12223072		71043529
	Chr5	173700114	Chr5	111463668	Chr4	139958906	Chr5	178662668	Chr4	106695711	Chr4	106695711
		180629495		112004635		140768929		180629495		108287723		108287723
	Chr6	27994809	Chr5	137419023	Chr5	44667765	Chr6	25857936	Chr5	77175855	Chr4	151320863
		29463498		139488889		49910087		26562433		77985321		153002089
	Chr6	40922113	Chr5	169858373	Chr5	71275894	Chr6	37818145	Chr5	131589914	Chr5	86441204
		41469562		170969444		73555480		40440789		131867517		88929211
	Chr6	47842726	Chr5	171511489	Chr5	111463668	Chr6	70642371	Chr6	28320468	Chr5	131589914
		51937898		172109066		112004635		71083045		28738772		131906947
	Chr6	54527706	Chr5	178662668	Chr5	137872173	Chr6	80585513	Chr6	30561692	Chr5	140336601
		55195369		180629495		139153625		82647587		31544026		141136483
	Chr6	105243262	Chr6	10520174	Chr5	168136033	Chr6	115102885	Chr6	44687767	Chr5	146464178
		120169406		11089829		178095202		116244328		45513755		165025942
	Chr6	152633125	Chr6	22374148	Chr5	178223874	Chr6	152633125	Chr6	54834286	Chr6	25928418
		153215692		22814614		180629495		153215692		55824119		26256290
	Chr7	17694780	Chr6	27436556	Chr6	3830151	Chr7	34083661	Chr6	77468918	Chr6	26449009
		18370203		28156603		12366280		36088555		79166984		29384159
	Chr7	93769310	Chr6	51865456	Chr6	27994809	Chr7	41312013	Chr6	133326866	Chr6	29463498
		94327034		52703024		29463498		42155063		134184353		30431602
	Chr8	647650	Chr6	107412431	Chr6	40922113	Chr7	73048344	Chr7	13630437	Chr6	38630800
		1151889		139354232		41469562		75205183		15037328		40367349
	Chr8	22280259	Chr6	152633125	Chr6	47842726	Chr7	155664126	Chr7	89464547	Chr6	138117136
		23607616		153215692		51937898		156169825		90512128		139174533
	Chr9	38762575	Chr7	11594993	Chr6	52111723	Chr8	3140500	Chr7	93859411	Chr7	3037641
		80571183		13647221		53544833		3366306		94025358		3463257
	Chr9	111831005	Chr7	34083661	Chr6	115102885	Chr8	112090806	Chr7	98771625	Chr7	13854270
		112533556		36088555		116244328		113712464		99203387		15037328
	Chr9	116767211	Chr7	41312013	Chr6	152633125	Chr8	115087798	Chr7	117751162	Chr7	65087151
		131087461		42155063		153215692		116178332		119057622		66538963
	Chr11	37902407	Chr7	73048344	Chr7	5187201	Chr9	19593778	Chr7	125563862	Chr7	81376443
		38748283		75205183		11347983		25797383		126363049		128572244
	Chr11	40383440	Chr7	93769310	Chr7	50776579	Chr9	26190057	Chr8	5556868	Chr7	155733549
		41677384		94327034		52932865		26648407		5944053		158812247
	Chr11	87107193	Chr7	155664126	Chr7	155664126	Chr10	88087	Chr8	52381624	Chr8	73202613
		87669641		156169825		156169825		6429678		53153482		81014928
	Chr11	89682187	Chr8	647650	Chr8	496821	Chr10	11997475	Chr8	56762102	Chr8	141843464
		90360718		1151889		2458818		35923553		57595407		143257639

	Chr11	120928850	Chr8	19574683	Chr8	5579321	Chr10	47164728	Chr8	84182673	Chr9	16828211
		132537965		20032491		5840921		51119344		85079809		17617988
	Chr11	133922969	Chr8	35434670	Chr9	111831005	Chr10	51259453	Chr9	16828211	Chr9	33101348
		134445626		37113474		112533556		53111013		17617988		38376351
	Chr12	21277285	Chr9	38762575	Chr9	116767211	Chr10	61095195	Chr9	33096619	Chr9	74419068
		23916968		80571183		131297604		61563904		36814086		74911269
	Chr12	39691763	Chr9	111831005	Chr9	138660206	Chr10	85922227	Chr9	73590943	Chr9	122370634
		40283954		112533556		139324980		86496172		76241584		123112473
	Chr12	118423916	Chr9	116767211	Chr10	4801703	Chr11	15401081	Chr9	116354595	Chr10	1311441
		119232296		140186312		5768180		16987351		123727267		1755025
	Chr13	31671559	Chr10	4055704	Chr10	85922227	Chr11	41556543	Chr10	24734721	Chr10	21808932
		32408197		4547508		86496172		42385965		26888212		23522641
	Chr13	35365044	Chr10	24383069	Chr11	15401081	Chr11	60593095	Chr10	47164728	Chr10	49371390
		36011511		42559839		16874445		61372588		50512270		66313209
	Chr13	48398174	Chr10	74416452	Chr11	37902407	Chr11	73832768	Chr10	94504955	Chr10	87399680
		49650816		76493170		38748283		74992435		94836083		91284204
	Chr13	81979755	Chr10	84670479	Chr11	40383440	Chr11	112700805	Chr10	116655867	Chr10	91767474
		82735155		85562477		41246340		113088180		117712802		92585046
	Chr13	88261590	Chr11	1822136	Chr11	47316646	Chr12	7409689	Chr11	5170989	Chr10	106367695
		107919211		6247696		48879101		34228293		5279313		107669590
	Chr13	113052302	Chr11	37902407	Chr11	80891332	Chr12	36785447	Chr11	22438325	Chr10	127324282
		114121631		38748283		95620677		51093857		69319975		127810968
	Chr15	64235510	Chr11	40383440	Chr11	112700805	Chr12	96494269	Chr11	110684600	Chr11	27735168
		76334504		41143665		113088180		97258308		111972700		30485520
	Chr15	82914976	Chr11	93214886	Chr12	19784513	Chr12	111852278	Chr11	113829236	Chr11	37178934
		83869754		96500421		32537115		113300743		117280039		40106260
	Chr15	88429541	Chr11	110684600	Chr12	39691763	Chr12	118423916	Chr11	122602596	Chr11	54691139
		89111157		111972700		40283954		119232296		133715739		56126682
	Chr16	30300387	Chr11	129270840	Chr12	96738707	Chr13	20698228	Chr13	17956717	Chr11	73832768
		31269424		132537965		101542704		36344108		19690184		74992435
	Chr16	51494327	Chr11	133922969	Chr13	27792206	Chr13	36895401	Chr13	22703627	Chr11	112700805
		52562664		134445626		36231087		37260794		23526272		113088180
	Chr16	83115664	Chr12	21277285	Chr13	58247078	Chr13	46487583	Chr13	52959314	Chr11	122985408
		84581605		23916968		60491341		51241636		53727005		133715739
	Chr17	18838397	Chr12	39691763	Chr13	61453476	Chr13	60165177	Chr14	40675306	Chr12	24438298
		19730045		40283954		78388053		78388053		41397854		24877007
	Chr17	45587906	Chr12	77727504	Chr13	85752319	Chr13	100389832	Chr14	54303168	Chr12	108623883
		45630865		101542704		107789005		105281282		55431054		109649498
	Chr17	53356034	Chr13	34263356	Chr14	70422209	Chr14	58660512	Chr15	23906668	Chr12	118118995
		74110845		35329698		71402949		60745462		30108777		119093708
	Chr18	58991846	Chr13	113052302	Chr15	93039223	Chr15	93039223	Chr16	65744712	Chr13	36895401
		76116152		114121631		93671981		93671981		67396803		37260794
	Chr19	1747166	Chr14	55450004	Chr15	95730163	Chr15	95730163	Chr17	13954919	Chr13	109667434
		2766372		56022394		96744124		96744124		22898583		110410132

	Chr19 62475189 63372500	Chr14 58660512 60745462	Chr16 2066490 2078253	Chr16 2066490 2078253	Chr17 45587183 45698385	Chr15 46157395 46670988
	Chr20 21502699 22273574	Chr14 102625250 103252751	Chr16 30300387 31269424	Chr16 23937673 27000638	Chr18 2905219 22555039	Chr15 78427042 78961599
	Chr21 19945837 26129974	Chr15 64235510 76334504	Chr16 51494327 55474446	Chr16 30818331 31765207	Chr18 22713449 22880828	Chr16 16090546 16174734
	Chr21 45770210 46367207	Chr15 82036073 83869754	Chr16 83115664 84581605	Chr16 52354558 56408106	Chr18 44303255 69183271	Chr16 81618912 83366401
	Chr22 14494244 16211813	Chr15 88429541 89111157	Chr17 6982970 7847211	Chr16 57042235 57602699	Chr19 1911031 8693319	Chr17 33729948 34594761
	Chr22 21061758 23555713	Chr16 62312305 64097923	Chr17 16266693 19377727	Chr17 2196654 9662875	Chr19 16008736 17562030	Chr17 45477321 45644854
		Chr16 65737699 67124288	Chr17 39720522 40356007	Chr17 45622227 45645685	Chr19 20051141 22345947	Chr18 10303183 22555039
		Chr17 16266693 19587273	Chr17 45622227 45644854	Chr17 58938359 74110845	Chr19 39777712 51203186	Chr18 22713449 22880828
		Chr17 45616365 45645685	Chr17 53356034 65242845	Chr18 7312666 28169259	Chr20 54550144 55660920	Chr18 44303255 50441081
		Chr17 63981737 74110845	Chr17 74116386 78072095	Chr18 58991846 76116152	Chr21 37254638 39311594	Chr18 70952589 76116152
		Chr18 7312666 37021893	Chr18 7415562 37021893	Chr19 1747166 2766372	Chr21 45770210 46396672	Chr19 1667865 8693319
		Chr18 51961012 59796144	Chr18 37413260 56937342	Chr19 60241570 61476840	Chr22 16111977 16707616	Chr20 8218102 44384258
		Chr18 60013901 70230155	Chr19 1747166 2664416	Chr20 9445388 9930745	Chr22 40052810 40852336	Chr20 49101662 54546170
		Chr19 1747166 2664416	Chr19 16596788 17562030	Chr20 21502699 22275414		Chr21 38175388 39311594
		Chr19 60241570 63788972	Chr19 41977210 43170120	Chr20 36005342 40165042		Chr21 45672710 46396672
		Chr20 13199844 40165042	Chr19 50575092 53202365	Chr21 45770210 46396672		
		Chr21 19944711 26129974	Chr19 60241570 63788972	Chr22 19274155 40852336		
		Chr22 14494244 16766273	Chr20 9445388 9930745	Chr22 40853155 48063820		
		Chr22 43309365 48073140	Chr20 21502699 22273574			
			Chr21 15375227 16972413			

Table

10.10 LRoHs identified from each affected individual using David Pike's method. The regions where the causal variant lies are labelled in **bold**. NB: This list may have been edited to ensure anonymity/confidentiality of the participants.

9.6.3. ARID causal/associated genes from GeneCards

No	Gene	Description	GeneCards ID
1	<i>ACTB</i>	actin, beta	GC07M005566
2	<i>AHI1</i>	Abelson helper integration site 1	GC06M135646
3	<i>AP4B1</i>	adaptor-related protein complex 4, beta 1 subunit	GC01M114437
4	<i>AP4M1</i>	adaptor-related protein complex 4, mu 1 subunit	GC07P099699
5	<i>AP4S1</i>	adaptor-related protein complex 4, sigma 1 subunit	GC14P031494
6	<i>APOE</i>	apolipoprotein E	GC19P045408
7	<i>ASPA</i>	aspartoacylase	GC17P003326
8	<i>ASPM</i>	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)	GC01M197053
9	<i>AVPR2</i>	arginine vasopressin receptor 2	GC0XP153167
10	<i>BBS2</i>	Bardet-Biedl syndrome 2	GC16M056569
11	<i>BBS4</i>	Bardet-Biedl syndrome 4	GC15P072978
12	<i>CC2D1A</i>	coiled-coil and C2 domain containing 1A	GC19P014016
13	<i>CC2D2A</i>	coiled-coil and C2 domain containing 2A	GC04P015471
14	<i>CEP290</i>	centrosomal protein 290kDa	GC12M088442
15	<i>CHKB</i>	choline kinase beta	GC22M051017
16	<i>CNTNAP2</i>	contactin associated protein-like 2	GC07P145863
17	<i>COL4A5</i>	collagen, type IV, alpha 5	GC0XP107683
18	<i>CRBN</i>	cereblon	GC03M003166

19	<i>DAG1</i>	dystroglycan 1 (dystrophin-associated glycoprotein 1)	GC03P049482
20	<i>DHCR7</i>	7-dehydrocholesterol reductase	GC11M071145
21	<i>DMD</i>	dystrophin	GC0XM031047
22	<i>DYM</i>	dymeclin	GC18M046570
23	<i>ELN</i>	elastin	GC07P073442
24	<i>FKRP</i>	fukutin related protein	GC19P047249
25	<i>FKTN</i>	fukutin	GC09P108320
26	<i>GNAS</i>	GNAS complex locus	GC20P057414
27	<i>GPHN</i>	gephyrin	GC14P066974
28	<i>KCNJ10</i>	potassium inwardly-rectifying channel, subfamily J, member 10	GC01M160007
29	<i>KIF1A</i>	kinesin family member 1A	GC02M241653
30	<i>L1CAM</i>	L1 cell adhesion molecule	GC0XM153126
31	<i>LAMA2</i>	laminin, alpha 2	GC06P129246
32	<i>MAN1B1</i>	mannosidase, alpha, class 1B, member 1	GC09P139981
33	<i>MAPT</i>	microtubule-associated protein tau	GC17P043971
34	<i>MCPH1</i>	microcephalin 1	GC08P006276
35	<i>MECP2</i>	methyl CpG binding protein 2 (Rett syndrome)	GC0XM153287
36	<i>MED23</i>	mediator complex subunit 23	GC06M131895
37	<i>MRT17</i>	mental retardation, non-syndromic, autosomal recessive, 17	GC04U901746
38	<i>MRT19</i>	mental retardation, non-syndromic, autosomal recessive, 19	GC18U900689
39	<i>MRT20</i>	mental retardation, non-syndromic, autosomal recessive, 20	GC00U931319

40	<i>MRT21</i>	mental retardation, non-syndromic, autosomal recessive, 21	GC00U931487
41	<i>MRT22</i>	mental retardation, non-syndromic, autosomal recessive, 22	GC00U931643
42	<i>MRT23</i>	mental retardation, non-syndromic, autosomal recessive, 23	GC11U901904
43	<i>MRT24</i>	mental retardation, non-syndromic, autosomal recessive, 24	GC06U902236
44	<i>MRT25</i>	mental retardation, non-syndromic, autosomal recessive, 25	GC12U901610
45	<i>MRT26</i>	mental retardation, non-syndromic, autosomal recessive, 26	GC00U931620
46	<i>MRT27</i>	mental retardation, non-syndromic, autosomal recessive, 27	GC15U901438
47	<i>MRT28</i>	mental retardation, non-syndromic, autosomal recessive, 28	GC06U902266
48	<i>MTHFR</i>	methylenetetrahydrofolate reductase (NAD(P)H)	GC01M011845
49	<i>NFKB1</i>	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	GC04P103422
50	<i>NR0B1</i>	nuclear receptor subfamily 0, group B, member 1	GC0XM030322
51	<i>NRXN1</i>	neurexin 1	GC02M050145
52	<i>NSUN2</i>	NOP2/Sun RNA methyltransferase family, member 2	GC05M006654
53	<i>OFD1</i>	oral-facial-digital syndrome 1	GC0XP013752
54	<i>OTC</i>	ornithine carbamoyltransferase	GC0XP038211
55	<i>PAH</i>	phenylalanine hydroxylase	GC12M103230
56	<i>PAX6</i>	paired box 6	GC11M031806
57	<i>POMT1</i>	protein-O-mannosyltransferase 1	GC09P134378
58	<i>PRSS12</i>	protease, serine, 12 (neurotrypsin, motopsin)	GC04M119201
59	<i>RAB3GAP1</i>	RAB3 GTPase activating protein subunit 1 (catalytic)	GC02P135809
60	<i>RB1</i>	retinoblastoma 1	GC13P048877

61	<i>RELN</i>	reelin	GC07M103112
62	<i>SMS</i>	spermine synthase	GC0XP021958
63	<i>SOD1</i>	superoxide dismutase 1, soluble	GC21P033031
64	<i>SOX3</i>	SRY (sex determining region Y)-box 3	GC0XM139585
65	<i>STS</i>	steroid sulfatase (microsomal), isozyme S	GC0XP007147
66	<i>TRAPPC9</i>	trafficking protein particle complex 9	GC08M140742
67	<i>TSC2</i>	tuberous sclerosis 2	GC16P002097
68	<i>TUSC3</i>	tumor suppressor candidate 3	GC08P015274
69	<i>UBR1</i>	ubiquitin protein ligase E3 component n-recognin 1	GC15M043235
70	<i>WT1</i>	Wilms tumor 1	GC11M032365
71	<i>ALX4</i>	ALX homeobox 4	GC11M044238
72	<i>CACNG2</i>	calcium channel, voltage-dependent, gamma subunit 2	GC22M036959
73	<i>CDH15</i>	cadherin 15, type 1, M-cadherin (myotubule)	GC16P089238
74	<i>DMD</i>	dystrophin	GC0XM031047
75	<i>DYNC1H1</i>	dynein, cytoplasmic 1, heavy chain 1	GC14P102430
76	<i>EPB41L1</i>	erythrocyte membrane protein band 4.1-like 1	GC20P034679
77	<i>FGFR2</i>	fibroblast growth factor receptor 2	GC10M123223
78	<i>GLI3</i>	GLI family zinc finger 3	GC07M041970
79	<i>GRIN1</i>	glutamate receptor, ionotropic, N-methyl D-aspartate 1	GC09P140032
80	<i>GRIN2B</i>	glutamate receptor, ionotropic, N-methyl D-aspartate 2B	GC12M013714
81	<i>HPRT1</i>	hypoxanthine phosphoribosyltransferase 1	GC0XP133594

82	<i>HRAS</i>	Harvey rat sarcoma viral oncogene homolog	GC11M000522
83	<i>KCNQ2</i>	potassium voltage-gated channel, KQT-like subfamily, member 2	GC20M062038
84	<i>KIRREL3</i>	kin of IRRE like 3 (<i>Drosophila</i>)	GC11M126326
85	<i>MBD5</i>	methyl-CpG binding domain protein 5	GC02P148778
86	<i>NAGLU</i>	N-acetylglucosaminidase, alpha	GC17P040687
87	<i>NDP</i>	Norrie disease (pseudoglioma)	GC0XM043808
88	<i>NF1</i>	neurofibromin 1	GC17P029421
89	<i>PACS1</i>	phosphofurin acidic cluster sorting protein 1	GC11P065837
90	<i>RBX1</i>	ring-box 1, E3 ubiquitin protein ligase	GC22P041347
91	<i>SCN8A</i>	sodium channel, voltage gated, type VIII, alpha subunit	GC12P051987
92	<i>SHH</i>	sonic hedgehog	GC07M155592
93	<i>SMARCA4</i>	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	GC19P011071
94	<i>SPAST</i>	spastin	GC02P032288
95	<i>SYNGAP1</i>	synaptic Ras GTPase activating protein 1	GC06P033387
96	<i>TSC1</i>	tuberous sclerosis 1	GC09M135766

Table 10.11 Examples of known ARID causal genes. The list was compiled from GeneCards using ‘mental retardation’ or ‘intellectual disability’ and ‘dominant’ or ‘recessive’ as the key words, source URL: <http://www.genecards.org/>.

9.6.4. Details on all ARID causal genes/mutations from literature

The table below (Table 10.12) lists all the known human ARID causal genes and the variants which cause them. The phenotypes caused as a result of the mutation, other information relating to the phenotype(s) and the references are also included below. The list was compiled with the help of Sian Lewis at the University of Bristol.

Gene name	Causal mutation	Location	Phenotype	Notes	References
<i>PRSS12</i> (neuronal serine protease neurotrypsin gene)	4-base pair deletion in exon 7. (ACGT) Deletion disrupted AatII site and caused a premature stop codon 147 nucleotides downstream of the deletion	Algerian	Non-syndromic Intellectual disability, IQ <50 Normal development until age 2 when intellectual disability became apparent	Immuno-electron microscopy on adult brain sections found that neurotrypsin is found the the presynaptic membrane and synaptic cleft Hybridisation experiments in human feral brains show that neurotrypsin is highly expressed brain regions associated with learning and memory Neurotrypsin-mediated proteolysis appear to be needed for normal synaptic function. (Needed for normal synaptic maturation) Therefore this mutation could be the bases of intellectual disability This mutation was not found in 200 unrelated controls individuals One affected individual in another Algerian consanguineous family from the same region	[526]
<i>CRBN</i> (cereblon)	C→T nonsense mutation at nucleotide position 1,274 of the LOC51185 Mutation causes an arginine residue to change to a stop codon resulting in a truncated protein Mutation occurs in exon 11 Found on chromosome 3 Mutation a N-myristoylation site removing casein kinase II	Religious immigrant community North America	Non-syndromic mild intellectual disability (IQ 50-70)	Gene codes for a an ATP dependent Lon protease Protein sequence conserved across species CRBN is highly expressed in the brain A member of the Lon-containing protein family is expressed in the human hippocampus, an important neuroanatomic region that is involved in long-term potentiation and learning. Lon proteins selectively degrade short-lived polypeptides and regulate mitochondrial replication and transcription.	[527]

	phosphorylation site at the C terminus			This mutation may prevent proper subcellular targeting and alter long-term potentiation in the hippocampus Paper suggest a role for ATP-dependent degradation of proteins in memory and. learning	
<i>CC2D1A</i>	Protein truncation The mutation is found on chromosome 19p13.12–p13.2 to an interval of 2.4 Mb deletion of 3589 nucleotides in intron 13 to 16 (exons 14 to 16 are deleted) Deletion causes removal of aa 408 to 547 causing a frame shift of the encoded protein immediately after the deletion, creating a nonsense peptide of 30 aa and a stop codon at position 438 of the mutant protein (G408fsX437)	Israeli Arab	Severe non-syndromic intellectual disability Psychomotor developmental delay during early childhood	CC2D1A function is unknown CC2D1A is a putative signal transducer participating in positive regulation of I- κ B kinase/NF κ B cascade CC2D1A is highly expressed in the cerebral cortex and hippocampus. CC2D1A gene contains two domains C2 and DM14 , C2 domain is found in proteins that are important in calcium dependent phospholipid binding while DM14 is unique to CC2D1A and its role is unknown but is repeated four times in human CC21A. The mutation removes one of DM14 domain and the C2 domains of the gene Mutation not seen in 300 controls. The expression patterns of the CC2D1A mRNA appear to be restricted mostly to the ventricular zone progenitors and neurons. In situ hybridisation in mice found that CC2D1A is expressed widely in the brain and strongly within the cerebral cortex and hippocampus. The paper suggests that a previously unknown signal pathway is important in human cognitive development.	[528]
<i>GRIK2</i> (also called <i>GLUK6</i>) Ionotropic glutamate receptor 6 gene	Single nonpolymorphic sequence change, a deletion of exons 7 and 8 (~120 kb) The mutation results in-frame deletion of 84 aa between amino acids 317-402	Iranian	Moderate to severe intellectual disability (Nonsyndromic)	GRIK2 encodes for a subunit of kainate receptors (KARs) which are highly expressed in the brain GRIK2 has been associated autism KARs are part of group of ionotropic receptors that are the targets of glutamate released in excitatory synapses.	[529]

	The mutation occurs the N-terminal region of GLUK6 the mutation comprises an inversion of ~80 kb, including exons 9, 10, and 11 in combination with a deletion of ~20 kb of intron 11			GRIK2 containing receptors are found in both pre and post synaptic cells in excitatory synapses suggesting that the mutation may effect local brain circuits Mutation or loss of NMDA/AMPA receptors resulting in severe derangement of neural function, but the role of ionotropic KARs in higher brain function has yet to be identified. The mutated gene product lacks the first ligand-binding domain, the adjacent transmembrane domain, and the putative pore loop, suggesting a complete loss of function of the GLUK6 protein making it unable to form a functional ion channel. None of 390 controls carried the mutation Paper suggest that GLUK6 is important for higher brain function in humans	
<i>TUSC3</i> (look at second article)	Deletion of 121,595bp between positions 15347852 and 15469447 on chromosome 8p22 The deletion includes the first exon of TUSC3 The mutation causes complete loss of TUSC3 function	Iranian French	Moderate to severe non-syndromic intellectual disability	TUSC3 is expressed widely throughout the human body including the brain. TUSC3 is assumed (as it similar to to the yeast Ost3 gene) to encode a subunit of the ER-bound oligosaccharyltransferase, this catalyses a step in the protein N-glycosylation process TUSC3 originally thought to be a tumour suppressor but as now of the patients have history of cancer this now seems unlikely. No further symptoms as loss of TUSC3 (other patients with congenital disorders of glycosylation experience further symptoms) partly compensated for by related gene on	[530] [531]

				Xq21.1 which encodes for the implantation-associated protein precursor. TUSC3 interacts with the alpha isoform of the catalytic subunit of protein phosphatase 1. Protein phosphatase 1 (PPPC1A) has been implicated in modulation of synaptic and structural plasticity, in mice was shown to impact on learning and memory. TUSC3-deficient patients exhibit MR as a result of loss of PPPC1A function. None of the 192 controls had the mutation.	
<i>TUSC3</i>	Deletion (170.673Kb) Deletion between 15521688 bp to 15692362 bp Mutation deletes the entire TUSC3 gene (except for the promoter and first exon) and its downstream region On chromosome 8p23	Pakistani	Severe non-syndromic intellectual disability	Mutation not found in 276 control Pakistani individuals.	[532]
<i>TRAPPC9</i>	Truncating mutation (R475X), in exon 7 of the gene TRAPPC9 A 4 bp deletion within exon 14 of TRAPPC9 was also identified in second family The mutation results in a significant degree of nonsense-mediated mRNA decay therefore any residual function in the truncated NIK- and IKK-b-binding protein (NIBP),	Pakistani	Moderate to severe Intellectual disability Mild cerebral white matter hypoplasia In some individuals microcephaly (abnormally small head) Individuals didn't learn to walk till around 5 years of age	TRAPPC9 encodes NIK- and IKK-b-binding protein (NIBP), which is involved in the NF-kB signalling pathway and directly interacts with IKK-b and MAP3K14 Individuals possessing the mutation expressed very little TRAPPC9 mRNA compared to controls. TRAPPC9 encodes the protein NIBP which is highly conserved across species. NIBP isoform 1 is expressed in the human brain where it is found in cell bodies and processes of neurones Knockdown of NIBP has been shown to reduce TNFα-induced NF-kB activation, prevent nerve growth factor-	[533]

	expression levels would be very low		Raised creatine phosphokinase levels However this is considered to be non-syndromic as not all these symptoms were seen in all patients	induced neuronal differentiation, and decrease Bcl-xL gene expression in PC12 cells These truncating mutations in these Pakistani and Iranian families may result in disruption of neuronal differentiation causing the cerebral white matter hyperplasia observed. A screen of 290 Pakistani controls for the mutation in exon 7 proved negative.	
	4 base pair deletion (c.2311–2314 delTGTT) Mutation causes in frame shift and premature truncation	Iranian	Non-syndromic intellectual disability		[534]
<i>TRAPPC9</i> Variant	Nonsense variant (c.1708C>T [p.R570X]) in exon 9 On chromosome 8q24	Tunisian	Intellectual disability Myelination defects Mild microcephaly Truncular obesity Hypertelorism mildly dysmorphic facial features	Mutation causes nonsense-mediated TRAPPC9 mRNA decay TRAPPC9 encodes a NF-κB-inducing kinase (NIK) and IκB kinase complex b (IKK-b) binding protein The paper suggest that NF-κB signalling is impaired in patients NF-κB transcription factor (TF) regulates the expression of a variety of genes and plays a key role in cellular processes such as innate and adaptive immunity, cellular proliferation, apoptosis, and development Studies show the role of neuronal NF-κB in memory and cognition and indicated that NF-κB activation is essential for long-term-memory formation, especially when hippocampus is involved Mutation not found in 1120 control chromosome including 196 of Tunisian origin	[535]
<i>TRAPPC9</i>	Nonsense mutation in exon 7 This mutation causes an arginine to change to a premature stop codon Two variants	Israeli Arab	Moderate to severe intellectual disability Microcephaly	Mutation causes loss of TRAPPC9 function TRAPPC9 has been linked to implicated in NF-κB activation and possibly in intracellular protein trafficking TRAPPC9 is highly expressed in the postmitotic neurons of the cerebral cortex	[537]

	Variant 1 c.1423C → T [p.R475X] Variant 2 c.1129C → T [p.R377X]			Patients with mutation in TRAPPC9 defects in axonal connectivity TRAPPC9 has higher expression in postmitotic neurones than progenitor cells Experiments with this gene in mice suggest that TRAPPC9 has an important role in the cytoplasm of postmitotic neurones This mutation was also observed in a Syrian family [536] due to close geographical and ethnical it is suggested that his is a founder mutation of this specific population and/or geographical area The mutation was not found in 153 control individuals	
<i>TRAPPC9</i>	Splice donor site mutation (c.1024+1G>T) t Mutation causes exon 3 (146 bp) to be skipped and exon 3 and 4 (275) This lead to the truncation of TRAPPC9	Pakistani	Intellectual disability Microcephaly	TRAPPC9 interacts withIKK-b and the NFkB-inducing kinase (NIK) Mutation in TRAPPC9 interferes with NFkB signaling in biological processes such as synaptic plasticity and neurogenesis resulting in ID Mutation not found in 100 Pakistani control alleles	[538]
<i>TRAPPC9</i>	Splice site mutation c.2851-2A>C before exon 18 Mutation leads to the skipping of exon 18and frame shift resulting in premature stop codon (p.T951Y fsX17)	Italian	Moderate to severe intellectual disability Peculiar facial appearance Obesity Hypotonia	Non-consanguineous parents Mutation lead to a loss of function of TRAPPC9 Paper suggest that loss of function mutations in this gene is associated with typical facial appearance, obesity, early onset, moderate-to-severe ID and highly specific brain abnormalities	[539]
Loci	4q26-4q28 (MRT17) 6q12-q15 (MRT18) 18p11 (MRT19) 16p12-q12 (MRT20) 11p15 (MRT21) 11p13-q14 (MRT23) 6p12 (MRT24) 12q13-q15 (MRT25)	Syrian	Non-specific intellectual disability	Heterozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci The linked regions vary in length between 1.2 and 45.6 Mb, and include between 9 and 625 RefSeq genes Apart from the TRAPPC9 mutation, none of the families with one linkage locus overlapped any of the 10 to-date-described ARID genes	[536]

	14q11-q12 (MRT26) 5q23-q26 (MRT27) 6q26-q27 (MRT28) Mutation was also identified in the TRAPPC9 (see Mochida et al., 2009)			Suggests that NS-ARID is very heretogeneous in Syiran populations	
<i>ZC3H14</i>	Mutation found on chromosome 14q31.3-q32.12 Homozygous nonsense mutation (R154X) in exon 6 Mutation disrupts ubiquitously expressed longer isoforms 1–3 but not the shorter brain- and testes-enriched isoform 4 of ZC3H14. Sequencing of ZC3H14 gene in second family showed on the same chromosome 14 locus a 25-bp deletion located 16 bp downstream of the 3'-end boundary of the annotated common exon 16 of ZC3H14.	Iranian	Intellectual disability	ZC3H14 encodes a conserved Cys3His tandem zinc finger polyadenosine RNA binding protein ZC3H14 mRNA transcripts can be found in the human central nervous system, Rodent ZC3H14 protein is expressed in hippocampal neurons and colocalizes with poly(A) RNA in neuronal cell bodies. A Drosophila melanogaster model with mutation of the gene encoding ZC3H14 ortholog dNab2, which also binds polyadenosine RNA showed that dNab2 is essential for development and required in neurons for normal locomotion and flight. Biochemical and genetic data show dNab2 restricts bulk poly(A) tail length in vivo, indicating this function may underlie its role in development and disease. The founding member of the encoded protein family is Saccharomyces cerevisiae Nab2. Nab2 is needed for viability and for proper 3'-end formation and poly(A) RNA export from the nucleus. Loss of Nab2 disrupts development and impairs neural function ZC3H14 is alternatively spliced to encode four ZC3H14 protein isoforms	[540]
<i>MED23</i>	Missense mutation (p. R617Q) This mutation doesn't affect expression protein stability,	Algerian	Non-syndromic Intellectual disability	MED23 is a subunit of the Mammalian Mediator complex, and a regulator of protein-coding gene expression by conveying information from transcription factors to the basal RNA polymerase	[541]

	architecture or composition of the Mediator complex.			II (Pol II) transcription machineries. Mediator senses developmental and environmental signals to ensure the correct result from the transcriptional machinery This gene is highly conserved across species Transcriptional dysregulation of these genes was also observed in cells derived from patients presenting with other neurological disorders linked to mutations in other Mediator subunits or proteins interacting with MED. MED23 gene that encodes one of the tail module's Mediator subunits. MED23 was identified as a suppressor of a hyperactive ras phenotype in <i>Caenorhabditis elegans</i> The mutation impaired the response of JUN and FOS immediate early genes (IEGs) to serum mitogens by altering the interaction between enhancer-bound transcription factors (TCF4 and ELK1, respectively) In M23/R617Q cells Pol II at the JUN promoter showed defective phosphorylation. The defective phosphorylation of Pol II reduced its ability to initiate transcription. JUN and FOS are implicated in learning and consolidation of a long-term memory trace Paper suggest that intellectual disability seen in these patients could be the result of problems with fine-tuning of IEG expression during development. This mutation was not found 608 control chromosomes.	
<i>TECR</i>	Pro182Leu mutation	Europe	Non-syndromic intellectual disability	Gene is found on chromosome 19p13 TECR (Trans-2,3-enoyl-CoA reductase) is a synaptic glycoprotein.	[542]

	Mutation causes substitution of leucine for a proline at amino acid 182 in TECR Proline is highly conserved in all species therefore it substituting leucine at amino acid 182 is likely to change protein function			TECR plays a role in the synthesis of very long-chain fatty acids in a reduction step of microsomal fatty acyl-elongation process Mouse ortholog of TECR is highly expressed in the nervous system The above suggest possible way mutation in TECR could disturb pathways resulting in NSID, however TECR could have an unknown function that affects communication between neurones or synaptic plasticity Screening within the European religious community found a carrier rate of 7.1% within that community. Evidence suggests a single ancestral founder haplotype Other nutation in TECR could be underlie related phenotypes in different populations such a schizophrenia and autism		
MAN1B1	Nonsense mutation c.1418G>A This mutation destroys a BamHI restriction site Mutation likely to cause a truncated protein missing residues required for substrate recognition.	Pakistani	Mild intellectual disability Dysmorphic features Delayed development	HBD region overlapped with that of MRT15, SNPs rs11103399– rs11137379; nucleotides 136,609,628–140,147,760 Neither of the Pakistani mutation were present in the 252 controls	MAN1B1 gene encodes the protein endoplasmic reticulum mannosyl oligosaccharide alpha 1,2-mannosidase (ERManI) ERManI and other class 1 a-mannosidases are members of the glycosyl hydrolase family 47 (GH47), are believed to be key enzymes involved in the maturation of N-glycans in the secretory pathway, and contribute to the timing and disposal of misfolded glycoproteins	[543]
	Missense mutation c.1189G>A Mutation creates a StyI restriction endonuclease cutting site	Pakistani	Non-syndromic intellectual disability (NS-ID) Delayed development Hyperphagia and overweight. Brain MRI showed small prominent perivascular space in the right parietal lobe, and cerebellar and	Found in three families from the same village, likely to have shared inheritance All three families shared a region of homozygosity by descent (HBD) on chromosome 9 between SNPs rs11103117 and		

			cerebral sulci that are mildly prominent but not enough so to constitute volume loss or atrophy	rs12238423, the locus was called MTR15. Haplotype comparison suggested a common founder for the three families despite familial connections not being found. As many of the traits could be familial traits this mutation was classed as non-syndromic. Neither of the Pakistani mutation were present in the 252 controls. Mutation affects the Glu397 residue which is located at the actgive pocket site where it interacts with glycan substates	through the endoplasmic-reticulum-associated degradation pathway. These mutations effect the enzymes ability to bind and recognise the oligosaccharide substrate. Paper suggests that the mutations disrupt the ER-associated degradation pathway	
	Missense mutationc.1000C>T One base pair deleted at Chr9:139,235,486 (hg19) in exon 7	Iranian	Mild to moderate non-syndromic intellectual disability (NS-ID)	This mutation was not observed in any of the 155 Iranian or 191 German controls. The mutation affects the Arg334 residue, which is believed to be located at the base of the active-site pocket.		
<i>LARP7</i>	Frameshift mutation On chromosome 4q24-q28.2	Saudi Arabian	Severe intellectual disability	Mutation leads to complete loss of LARP7 LARP7 is the chaperone of 7SK, it binds to 7SK protecting it from degradation.		[544]

	7bp duplication in exon 8 (c.1024_1030dupAAGGATA, p.T344Kfs*9), This frameshift causes premature truncation of the peptide after 9 residues		Facial dysmorphism (malar hypoplasia, deep-seated eyes, broad nose, short philtrum, and macrostomia) Primordial Dwarfism	Depletion of 7SK, an abundant cellular noncoding RNA, underling these symptoms Complete loss of LARP7 protein leads to 7SK depletion. LARP7 forms a complex with 7SK snRNA with other proteins which helps to sequester the general transcription factor P-TEFb into an inactive state stopping RNAPII from elongating transcripts. 7SK effects expression of a wide number of genes though its inhibitory effect on the positive transcription elongation gactor b(P-TEFb) 7SK also has a cometing role in HMGA1-mediated transcriptional regulation. Mutation was not found in a screen of 188 controls	
<i>ST3GAL3</i>	Chromosome one within MRT4 locus 2 missense mutations at the intervals 10.1Mbp and 9.3Mbp The two mutations shared region is 7.9Mbp The mutations effected positions in exon 2 (NM_006279.2: c.38C>A, p.Ala13Asp) and exon 14 (NM_006279.2: c.1108G>T, p.Asp370Tyr	Iranian	Non-syndromic intellectual disability	ST3GAL3 encoges the Golgi enzyme β -galactoside- α 2,3-sialyltransferase-III which in human FORMS THE SIALYL Lewis a epitope on protein These mutations cause ER retention of the Golgi enzyme and drastically impair ST3Gal-III function Glycocalyx is an essential information carrier in living systems. The mutation in exon 2 affects the transmembrane domain The mutation in exon 14 effects the catalytic domain These mutations were present in sufferers but not found in 1000 control chromosomes	[545]
<i>NSUN2</i>	3 different mutations Nonsense mutation c.679C>T [p.Gln227*] Nonsense mutation c.1114C>T [p.Gln372*] Both nonsense mutations cause loss of NSUN2 transcripts in homozygous individuals	Iranian Kurdish	Moderate to serve intellectual disability Facial dysmorphism (long face, characteristic eyebrows, a long nose, and a small chin)	Paper looked at 3 independent consanguineous families NSUN2 encodes a tRNA-methyl-transferase which catalyses the intron- dependent formation of 5-methylcytosine at C34 The fourRNA-methyl-transferase transcripts that NSUN2 encode are expressed in the brain at various developmental stages. All three mutations cause the loss of NSUN2 function	[546]

	The third mutation causes an intronic exchange of an adenine for cytosine 11 nucleotides before exon 6 g.6622224A>C [p.Ile179Argfs*192]) Intronic nucleotide change causes exon 6 to be skipped during slipping causing a change in reading frame resulting in a premature stop codon causing the loss of the main transcript in tissues All 3 at the MRT5 locus			In Drosophila model the deletion of the NSUN2 ortholog lead to severe short-term memory deficit. Found that NSUN2 not essential for survival but is crucial in higher cognitive functioning. NSUN2 expression observed in the foetal brain, it is possible that the phenotype of affected individuals is the result of proteomic shifts caused by the absence of NSUN2 at critical stages during brain development Indicates that RNA methylation plays important role in cognition. The paper suggest that NSUN2 may play a role in translational regulations needed for proper synaptic plasticity subsequently effecting memory It is also possible that the disease mechanism may impair methylation of hemimethylated DNA as that is also a target of NSUN2 activity. None of the mutations were observed in the 185 ethnically match controls or 540 chromosomes from German controls.	
NSUN2	Missense change c.2035G>A (p.Gly679Arg)	Pakistani	Intellectual disability Distal myopathy	The mutation occurs at a conserved residue within NSUN2 This gene encodes a methyltransferase that catalyzes formation of 5-methylcytosine at C34 of tRNA-leu(CAA) and plays a role in spindle assembly during mitosis as well as chromosome segregation. In mice brains NSUN2 is localised to the nucleolus of Purkinje cells in the cerebellum The mutation to arginine at this residue results in NSUN2 failing to localize within the nucleus	[547]
CC2D2A	Splice-donor-site mutation (IVS19+1:G/C) On chromosome 4 (4p15.2-p15.33) Mutation skips exon 19 causing a frame shift	Pakistani	Mild to moderate intellectual disability Retinitis pigmentosa (degenerative eye condition)	CC2D2A is suggested to have a role in calcium dependent signal transduction CC2D2A expression was found in a range of adult tissues. Human fetal brain Marathon Ready cDNA showed strong expression in the brain	[548]

	The mutation results in a truncated protein missing the C2 domain			C domain (protein kinase C conserved region 2) is a calcium dependent membrane –targeting module found in proteins involved in membrane trafficking or signal transduction C- terminal conserved across species Suggest similar function (or pathway) of CC2D2A to CC2D1A and therefore maybe important for neuronal development The mutation was not found in 460 control Pakistani controls.	
<i>TBC1D24</i>	Mutation within 16p13.3 Single missense change, TBC1D24 (c.751T>C, p.F251L)	Arab-Israeli	Mild to moderate intellectual disability Folcal epilepsy Mild dysarthria and ataxia Subtle cortical thickening	TBC1D23 is part of a large gene family encoding TCB domain proteins which are predicted to be Rab GTPase activators Intellectual disability associates with seizures in around 21% of cases Phenylalanine (F) at position p.251 is conserved in mammals TBC1D24 is expressed in mouse embryo hippocampal neurons, and the developing mouse brain Over-expression of the wild-type TBC1D24 protein in mouse embryo increased the length of primary axons, increased arborisation and ectopic axon specification was also observed, no change in over-expression of mutant protein. The above suggests that this mutation leads to a loss of TBC1D24 protein function TBC1D24 protein is a potent modulator of primary axonal arborization and specification in neuronal cells This mutation was not found in 210 control chromosomes from a matched Arab population	[549]
<i>ADAT3</i>	Missense mutation On chromosome 9	Arab	Intellectual disability Strabismus	ADAT3 encodes one of two eukaryotic proteins that are needed for the deamination of adenosine at position 34 to inosine in t-RNA Ancient ancestral haplotype found in all eight families	[446]

	Mutation causes a change from a valine residue to methionine (c.382G>A, p.V128M)			The eight families are from different geographical regions Estimated that the mutation occurred 1600 years ago Paper suggests human brain is sensitive to defects in the regulation of translation Mutation not seen in 194 ethnically matched controls	
<i>GMPPA</i>	Two variants: Nonsynonymous change c.19C>T (p.Pro7Ser) Nonsense mutation c.295C>T (p.Arg99*)	Pakistani	Intellectual disability Achalasia Alacrima, Delayed developmental milestones, Gait abnormalities	Guanosine diphosphate (GDP)-mannose pyrophosphorylase A (GMPPA) Symptoms similar to triple A syndrome GMPPA is a homolog of GMPPB GMPPB catalyses the formation of GDP-mannose GDP-mannose is an essential precursor of glycanmoieties of glycoproteins and glycolipids GDP-mannose pyrophosphorylase A activity is unchanged and GDP-mannose levels are increased in lymphoblast in patients with GMPPA mutation suggesting that it is a GMPPB regulatory subunit	[550]
<i>PIGT</i>	c.547A>C (p.Thr183Pro) Variant - TUBB1 NM_030773:c.925C>T:p.Arg309 Cys thought unlikely to cause patient symptoms as mutations in this gene cause macrothrombocytopenia which wasn't seen in the patients		Severe intellectual disability Severe motor disability Hypotonia Seizures Abnormal skeletal, renal, endocrine and ophthalmologic Mild dysmorphic facial features	PIGT highly conserved throughout evolution. PIGT encodes phosphatidylinositol-glycan biosynthesis class T (PIG-T) protein, which is a subunit of the transamidase complex that catalyses the attachment of proteins to GPI. GPI acts as an anchor in the plasma membrane for extracellular proteins This mutation causes the impairment in membrane anchoring of GPI In vivo assay on zebrafish embryos suggest that this mutation impairs enzyme function Mutation causes partial loss of function of the protein Mutation not found in 200 Danish exomes or 100 local sample exomes.	[551]
<i>AP4E1</i>	Nonsense mutation (p.R1105X)	Moroccan	Intellectual disability Progressive spastic paraplegia	The mutation doesn't affect AP4E1 mRNA levels however components of the AP-4 (adapter protein-4) complex are affected.	[552]

			Short stature Two syndromes – hereditary spastic paraplegia (HSP) and mycobacterial disease	AP-4 is made up of four subunits	
<i>WDR62</i>	Deletion mutation c.1143delA In exon 9, On chromosome 19 This causes a frame shift resulting the protein being truncated (p.H381PfsX48)	Pakistani	Intellectual disability Primary microcephaly Schizencephaly Hypoplasia of corpus callosum	WDR62 encodes a nuclear localized spindle pole protein that is essential for proper neuroprogenitor cell division The truncated WDR62 protein is missing six WD40 repeat domains in addition to the C terminal, which contains several phosphorylation sites Mutation results in loss of protein function causing loss spindle poles of the dividing neural progenitor cells during mitosis effecting brain development. Mutation not found in 100 normal Pakistani individuals	[553]
<i>ERLIN2</i>	Two nucleotide insertion (c.812_813insAC) This insertion leads to a frame shift causing a premature stop codon (p.Asn272ProfsX4) This mutation causes the protein to truncate by about 20%	Turkey	Severe intellectual disability Motor dysfunction Multiple progressive joint contractures Large vacuoles containing flocculent material was found in leukocytes	ERLIN2 encodes endoplasmic reticulum (ER) lipid raft- associated protein 2 that mediates the ER-associated degradation of activated inositol 1,4,5-trisphosphate receptors (ITPRs) and other substrates ITPRs are very important in calcium homeostasis as ITP regulates calcium release from the ER via ITPRs Cytoplasmic free Ca ²⁺ concentration is a signaling system involved in the regulation of numerous cellular processes, including motor learning and memory, muscle contraction, etc The truncation of the protein leads to the loss of domains in the C terminal that mediate the association with detergent- resistant membranes and the oligomerization into either homo-oligomers of ERLIN2 or hetero-oligomers The large vacuoles in patient white blood cells indicated that the cellular pathology could be ER dysfunction This gene is highly conserved across species The mutation was not found in 109 controls	[554]

<i>ANKH</i>	Missense mutation L244S (c.731T->C) in exon 7 at 5p15 locus	Turkish	Intellectual disability Deafness Ankylosis Mild hypophosphatemia Progressive spondylarthropathy Osteopenia Painful small joint soft-tissue calcifications	Gene product ANK regulates tissue mineralization by transporting pyrophosphate to the extracellular space ANK is broadly expressed in the human body and is found in neuronal cell bodies and dendrites L244 is a highly conserved amino acid Ankylosis leads to hearing loss when it effects the hearing ossicles Mutation leads to loss of function despite normal ANK protein expression in the plasma membrane The heterozygous individuals demonstrated late-onset arthrosis from the third decade of life without ankylosis and mental retardation Mutation not found in 218 Turkish and 200 Caucasian control alleles	[555]
<i>ASPM</i>	A maternally inherited (c.2389C > T) in exon 6 this mutation is predicted a p.Arg797X nonsense mutation A two base pair deletion c.7781_7782delAG, in exon 18 was paternally inherited This mutation causes a frame shift resulting in a premature stop codon (p.Gln2594fsX6). At the MCPH5 locus	Algerian	Mild to severe intellectual disability Severe microcephaly Simplified cortical gyration Low to normal birth weight Severe hypoplasia of the frontal lobes and moderate posterior parietal atrophy, an anterior orientation of the insula bilaterally and a thick corpus callosum was observed by MRI	ASPM mutations are the commonest cause of autosomal recessive primary microcephaly (MCPH) More than 30 distinct ASPM mutations have been reported, including nonsense, frameshift and splice site mutations, a single missense mutation and a translocation breakpoint. ASPM highly expressed in fetal brain where its gene product maintains symmetric proliferative divisions. These mutations predict a loss of functions by either nonsense-mediated mRNA decay or synthesis of a severely truncated inactive ASPM protein These mutations were not found in 50 control individuals	[556]
<i>ASPM</i> (abnormal, spindle-like, microcephaly-associated)	19 mutations (see Table 1 in link: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1180496/table/TB3/)	Pakistani Dutch	Brian fails to develop correctly This causes microcephaly	Mutations in this gene are thought to be the most common cause of human autosomal primary microcephaly (MCPH) Comprehensive screen of the gene in 23 consanguineous families	[557]

	All mutations predicted to lead to truncation of the protein Two further mutations discovered 3663delG in a northern Pakistani family and 1365G-T in a Turkish family	Jordanian Saudi Arabian Yemeni	Intellectual disability	In MCPH the brain is small but structurally normal with no other neurological defects than intellectual disability The predicted ASPM protein is conserved Mutations in the drosophila ortholog asp, are embryologically lethal In drosophila asp is associated with the minus ends of microtubules at spindle poles in mitosis and meiosis In the CNS a large number of neurone progenitor cells are unable to complete asymmetric cell division stopping during metaphase. Mutations found throughout gene Papper suggest that truncation of ASPM protein is a major cause of MCPH	
<i>COH1</i>	Novel deletion spanning exon 6 to exon 16 (c.3348_3349delCT) On chromosome 8q22-q23 Mutation causes severe truncation of the protein	Greek effected population from two small neighbouring Greek islands	Cohen syndrome: Intellectual disability , microcephaly, growth delay, severe myopia, Progressive chorioretinal dystrophy, facial anomalies, slender limbs with narrow hands and feet, tapered fingers, short stature, kyphosis and/or scoliosis, pectus carinatum, joint hypermobility, pes calcaneovalgus, and, variably, truncal obesity	Cohen syndrome is reported worldwide but is overrepresented in Finland Difference between finnish patients and those examined in this study were observed. Paper suggests the founder effect in the spread of this mutation. COH1 encodes a putative transmembrane protein Population of these islands are decedents of a small number of families that settled on the islands during the 18th century In situ hybridization studies in mice showed abundant Coh1 gene expression in the central nervous system of juvenile and adult animals, but not in embryos Indications that the mutation in COH1 interfere with the function of postmitotic neurones by possibly disrupting dendritic or axonal outgrowth this could be consistent with the CS neurological profile	[558]
<i>MCPH1</i>	Deletion (150-200 kp)	Iranian	Intellectual disability	Indications suggest that the mutation causes a loss of function of the gene	[559]

	On the short arm of chromosome 8 Deletion includes first 6 exons and promoter of the gene		Borderline to mild microcephaly	MCPH1 gene product microcephalin is expressed in fetal brain, liver and kidney and at lower levels in adult tissues. In situ hybridisation has shown in mice the MCPH1 gene is expressed in the developing forebrain during neurogenesis. Experimental evidence suggests that MCPH1 has a role in DNA repair and is required to prevent premature onset of chromosome condensation. It appears that MCPH1 has a role in the control of cell cycle timing. Impairment of the gene could lead to primary microcephaly and intellectual disability by interfering with neurogenic mitosis in the developing brain	
<i>SYNGAP1</i>	Three de novo truncating mutations identified K138X (nonsense mutation) - Causes a truncated protein lacking the RASGAP domain R579X (nonsense mutation) -Truncated protein in middle of the RASGAP domain L813RfsX22 (c.2438delT) -Truncate protein just after the RASGAP domain		Non-syndromic moderate to severe intellectual disability	Dominant autosomal inheritance Non-consanguineous families Mutation absent in patient's parents SYNGAP1 encodes ras GTPase-activating which is crucial for cognition and synapse function SYNGAP1 is selectively expressed in the brain GTPase-activating protein is a component of the NMDA-receptor complex which acts downstream of the receptor blocking insertion of the AMPA receptor at the postsynaptic membrane by inhibition of the RAS-ERK pathway In mice with heterozygous mutation have impaired synaptic plasticity and learning, homozygous mice die shortly after birth (indicates important role of SYNGAP1 in postnatal development) RASGAP domain activates ras GTPase These mutations result in the production of proteins that lack domains, such as RASGAP and QTRV, that have been shown to be important for the synaptic plasticity and spine morphogenesis that are required for learning and memory	[560]

				None of these mutations were found in 142 individuals with autism spectrum disorders, 143 with schizophrenia and 190 control individuals Heterozygous missense variant (I1115T), a benign polymorphism, was present in patients with non-syndromic mental retardation and control subjects Three unique heterozygous missense variants were found in subjects with autism spectrum disorders (P1238L) and schizophrenia (T1310M and T790N)	
<i>C2orf37</i>	Single base pair deletion (c.436 delC) In exon 4 Mutation causes a frameshift (p.Ala147HisfsX9) in the β-isoform The mutation was found in six additional Saudi families but not in 274 Saudi control chromosomes	Saudi	Woodhouse–Sakati Syndrome (WSS): Hypogonadism, alopecia, intellectual disability, deafness, diabetes mellitus and progressive extrapyramidal defects	The protein encoded by C2orf37 appears to have a role in the nucleolus suggesting that a nucleolar defect underlies WSS The Saudi mutation is a founder mutation that arise and estimated 55 generations ago Further 3 loss of function mutations found in other ethnicities Gene has extreme splicing variability found 30 isoforms This mutation is unlikely to be pathogenic for every splice form The amino acid sequence from this gene is highly conserved across species Nucleolar expression in mouse embryos especially in the brain, liver and skin. This expression is consistent with WSS symptoms All mutations cause truncation of C2orf37	[561]
	1pb deletion (c.50 delC) also causes a frameshift in the β-isoform This mutation was not found in 210 control chromosomes	Eastern european			
	Splice donor sites c.1422+5G > T Mutation causes a big drop in splicing efficiency This mutation was not found in 196 control chromosomes	Indian			
	Splice donor mutation c.1091+6T > G Mutation causes the skipping of exon 10 leading to a frameshift	Middle eastern			

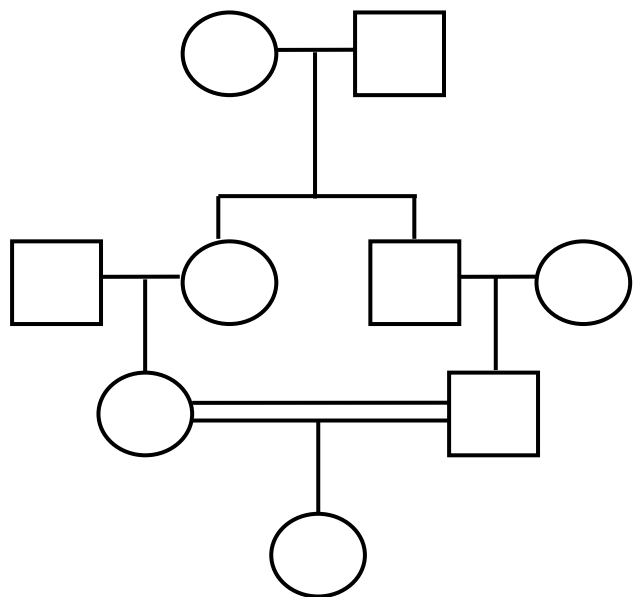
	This mutation was not found in 250 control chromosomes				
<i>C2orf37</i>	c.127-3delTAGinsAA Replaces the last 3 base pairs of intron 1 disrupting the splice acceptor site of exon 2 causing aberrant splicing	Turkish	Woodhouse-Sakati Syndrome (WSS)	All mutation result in the truncation of C2orf37 No correlation between the length of the mutated protein and the severity of the phenotype suggesting that intact protein needed for it to be functional Therefore all known mutations are likely to cause loss of function in the protein rather than nonsense mediated mRNA decay Intrafamilial variability was observed suggesting modifiers have an important role of modifiers	[562]
	Nonsense base substitution c.341C>A in exon 4 This mutation causes a premature stop codon	Italian			
	Base substitution c.387G>A in exon 4 causing a Trp-to-Ter change(p.W129X)	French gypsy			
	Nonsense mutation c.906G>A in exon 9 This causes a premature stop codon downstream of the mutation	Italian Non-consanguineous			
<i>C2orf37</i>	Splice mutation (c.321+ 1 G>A) in intron 3 Mutation result in a dramatic drop in splice site recognition leading to abolition of the normal splice donor site	Pakistani	Woodhouse-Sakati Syndrome (WSS)	This mutation also leads to protein truncation The changes of the splice donor site could lead to loss of function of the C2orf37 protein by nonsense mediated mRNA decay or synthesis of truncated protein Mutation not found in 100 ethnically matched control individuals	[563]
<i>C2orf37</i>	One base pair deletion (c.436delC) in exon 4	Qatari (Bedouin tribe)	Woodhouse-Sakati Syndrome Milder phenotype of WSS lack evidence of diabetes mellitus and extrapyramidal symptoms	The variable expression seen might suggest the presence of disease modifying factors In this case intellectual disability is present in early life but other symptoms may not occur until adolescence	[564]

<i>NDE1</i>	Deletion of exon 2 (4,296 nucleotides) c.-43-3548_83+622del Mutation removes mRNA 43 nucleotides upstream of the initiation codon Mutation results in a null allele	Turkish	Microhydranencephaly Lissencephaly or microlissencephaly Microcephaly, motor and mental retardation Brain malformations that include gross dilation of the ventricles with complete absence of the cerebral hemispheres or severe delay in their development	Paper reports mutations found by others in Saudi Arabian, Pakistani and Turkish families – 2 frame-shifts c.684_685del (p.Pro229TrpfsX85) and c.773dup (p.Leu245ProfsX70) and a splice mutation (c.81+1G>T or p.Ala29GlnfsX114) Protein product of NDE1 plays a role in mitosis C-terminal of the protein is essential for the localization of the protein to the centrosome and for its interaction with CENPF and dynein Ablation of Nde1 in mouse produced a brain one third smaller in size, with the most dramatic reduction affecting the cerebral cortex Mutation not found in 109 control individual;s from the population	[565]
<i>SWIP</i> [Strumpellin and <i>WASH</i> (Wiskott-Aldrich syndrome protein and scar homolog)-interacting protein]	Missense mutation, c.3056C>G in exon 29 of kiaa1033 this mutation causes Pro1019Arg exchange in SWIP Variant: c.1085A . T (p.His362Leu) in exon 5 of LHX5, an LIM homeobox gene	Iranian	Non-syndromic intellectual disability	The LHX5 gene is expressed during embryonic development and is involved in the anatomic formation of the hippocampus however as no malformation of the hippocampus was observed variant LHX5 appears to be non pathogenic SWIP is a subunit of the WASH complex WASH complex controls the polymerization of actin at the surface of endosomes WASH-dependent actin polymerization promotes scission of transport intermediates of the different endosomal routes and affects recycling, degradation and retrograde pathways Mutation PRO1019ARG destabilizes the SWIP and produces lower levels of strumpellin and WASH Deleterious effect on WASH complex assembly leads to severe neuronal malfunction, either through interaction with the Rho pathway or by affecting other endosomal pathways LHX5 and KIAA1033 both could contribute to the described phenotype Mutations not found in 1000-genome database and 200-Danish exome data	[566]

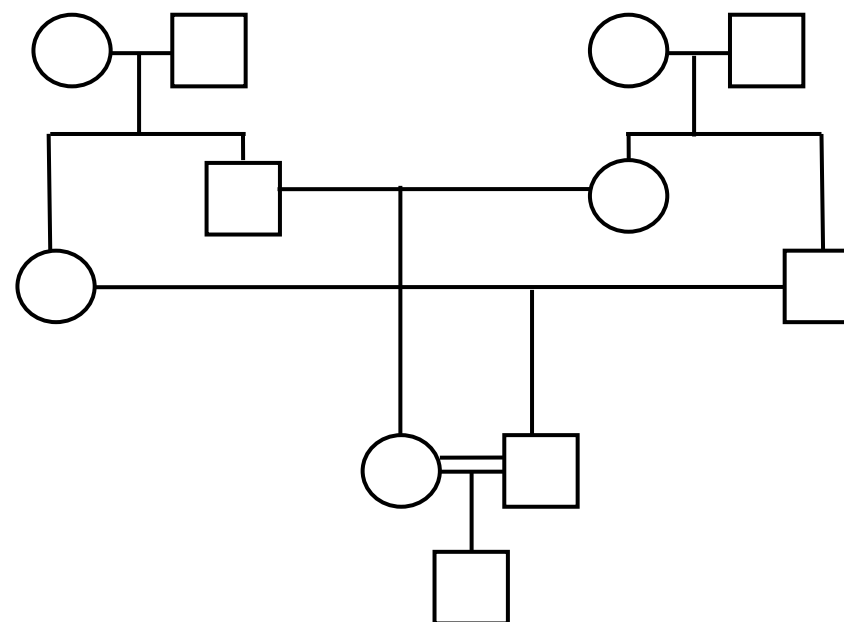
<i>FLJ90130</i>	Point mutation in exon 2 (C->G) that predicted a Y16X change Deletion of last nucleotide of exon 8	Dominican republic	Intellectual disability Mental Retardation and Abnormal Skeletal Development Dyggve-Melchior-Clausen Dysplasia (DMC)	This gene is highly conserved Mutations causes early truncation and loss of function in FLJ90130 or mRNA instability and no protein Transversion in exon 5 (T->A) causing a Y132X change and a point mutation that causes a N469Y missense change were found in a child that was a compound heterozygote. Unable to be sure that this is pathological, however evidence suggest that it is likely that the N469Y mutation is likely to be FLJ90130 is expressed in many tissues FLJ90130 is suggested to be a transmembrane protein	[567]
Region	De novo 17q21.33 microdeletion on chromosome 17 1.8Mb in size 24 RefSeq genes in this deletion include CA10 and CACNA1G genes	Lithuanian (non-consanguineous)	Mild intellectual disability Growth retardation Microcephaly Long face, large beaked nose, thick lower lip, micrognathia and other dysmorphic features	CA10 gene is highly expressed in the brain and plays an important role in CNS and brain development CACNA1G gene encodes voltage-dependent a calcium channel which is involved in neuronal oscillations and resonance in pacemaking activity in central neurones and neurotransmission CHAD gene is a candidate for the prenatal and postnatal growth retardation as its product chondroadherin is a cartilage protein with cell binding properties CHAD, CA10 and CACNA1G gene may be responsible for the overall phenotype by haploinsufficiency Deletion encompasses 24 known genes, 21 protein coding and 3 non-coding RNA genes Seventeen of the known genes have been shown to be expressed in brain 3 individuals with informative overlapping deletions were also found	[568]

Table 10.12 All known human ARID causal genes and the variants which cause them. The respective references are also included.

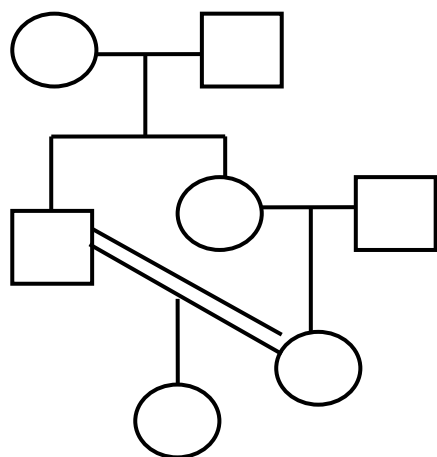
9.7. Consanguineous Unions



Union between **first cousins**, a
very common form of
consanguineous unions
 $F = 0.0625$

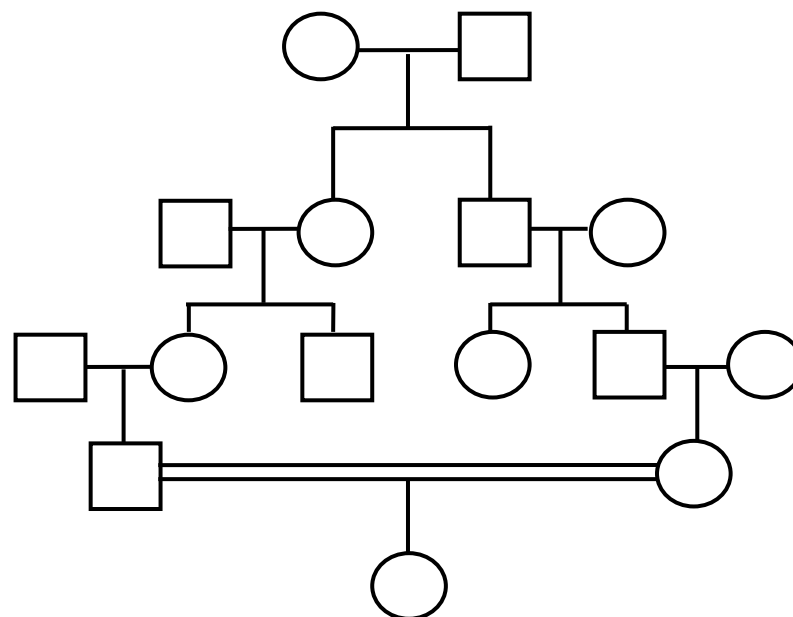


Union of **double (first) cousins**,
where both parents share the same
four grandparents
 $F = 0.125$



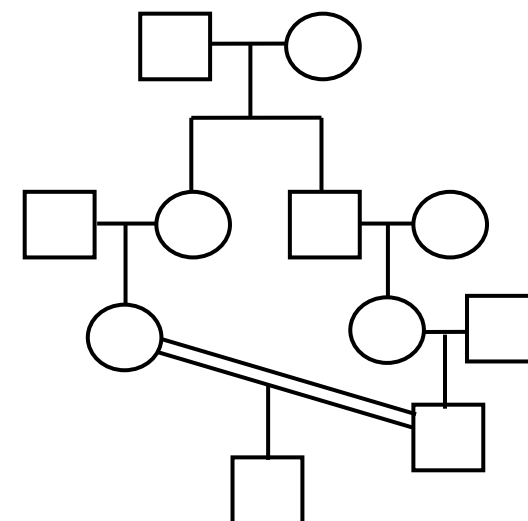
Uncle-niece unions are common in Southern India, an uncommon type anywhere else (Aunt-nephew unions is also consanguineous, and their offspring would be expected to have similar F value)

$$F = 0.125$$



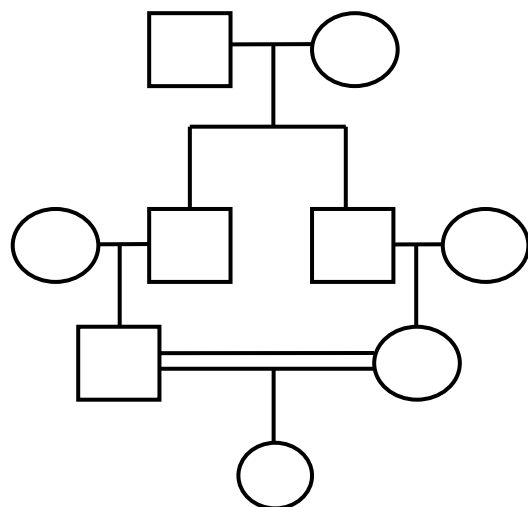
Unions between **second cousins** occur when offspring of first cousins marry, union with lowest level of kinship which is considered 'consanguineous'

$$F = 0.015625$$

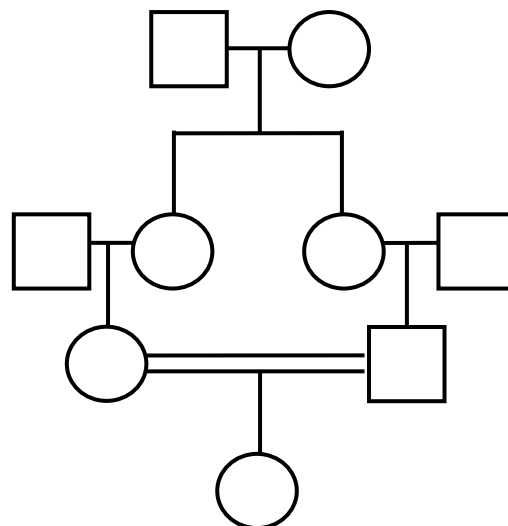


Unions between **first cousins once removed**

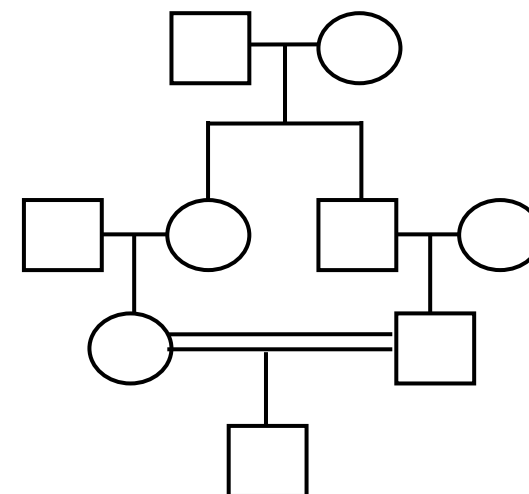
$$F = 0.03125$$



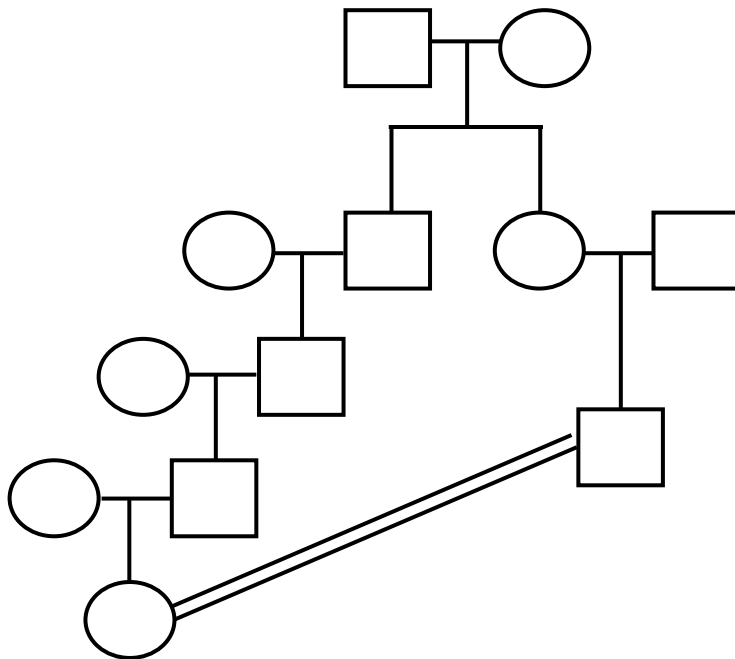
Unions between **patrilineal parallel first cousins** (father's brother's daughter) occur when offspring of first cousins marry
 $F = 0.0625$



Unions between **matrilineal parallel first cousins** (mother's sister's daughter) occur when offspring of first cousins marry
 $F = 0.0625$

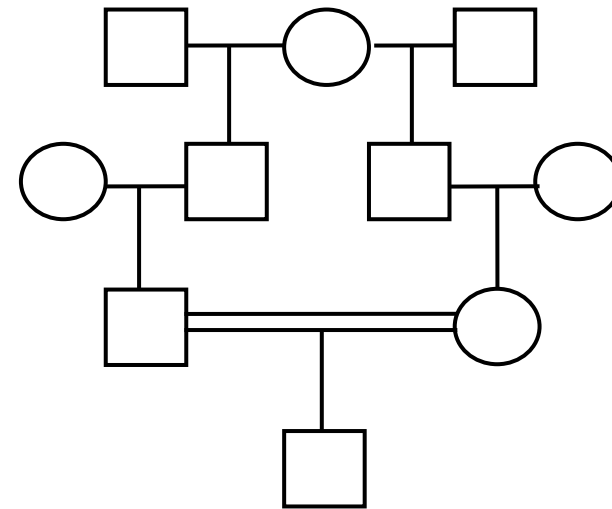


Unions between **cross first cousins** (mother's brother's daughter or father's sister's son) occur when offspring of first cousins marry
 $F = 0.0625$



Unions between **first cousins twice removed** would also be considered consanguineous, however they very rarely, if ever, occur

$F = 0.015625$



Unions between **half cousins**
very rarely occur due to the
obvious statistical reasons
 $F = 0.03125$

9.8. Other scripts, Files and UNIX commands used

Below are the Python and Perl scripts used in this thesis which have been written either by me or through collaboration with others (where latter is the case, the contributors were stated). Also included in this section are the UNIX commands used to use specific features of UNIX to serve a certain purpose, which is stated at the top of each section.

Annotating VCF files using VEP*

- 1- Download latest package (and **plugins) from Ensembl website: (www.ensembl.org/info/docs/variation/vep/index.html)
- 2- tar xvf downloaded file(s)
- 3- perl INSTALL.pl - and download *Homo sapiens* cache(s)
- 4- perl variant_effect_predictor.pl -i **file.vcf** -o **file.vep** --protein --cache --regulatory --gmaf --force_overwrite --sift b --polyphen b --plugin Condrel,/data/home/user/to/config/Condrel/config,b --plugin Conservation,GERP_CONSERVATION_SCORE,mammals --fork 8 --canonical

*includes SIFT, Polyphen and Condrel predictions in the output (see for other options: http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html)

**For example, before using the Condrel plugin:

- 1- Download latest Ensembl plugins from: https://github.com/ensembl-variation/VEP_plugins
- 2- tar xvf downloaded file
- 3- mv Condrel.pm ~/.vep/Plugins (create Plugins folder if not there; also .vep is a hidden folder)
- 4- edit the condrel_SP.conf file (in config/Condrel/config/) and set the 'condrel.dir' parameter to
/data/home/user/to/variant_effect_predictor/ensembl-variation-VEP_plugins-e6cec6a/config/Condrel

Using FATHMM (using output from step 4)

- 1- grep missense **file.vep** | grep CANONICAL > **file.vep.missense**

- 2- python **missense2fathmm.py** (converts all .missense files to FATHMM input format)
- 3- Copy and paste to 'Protein submission' in FATHMM (Disease Ontology)
- 4- Download .tar file (end of page) and unzip using: tar xvf **.tar file**

To convert plink files to vcf files:

```
python convert_plink_to_vcf.py <ped file> <map file> <UCSC human genome in 2bit
format>
```

then

```
grep -v N input_file > output_file
```

then

```
add #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT N1
```

To convert 23andme data to plink files (example):

```
perl convert_23andme_to_plink.pl <23andme.txt> 2 1 1 1 2 2
```

To convert vcf files to BEAGLE format (and vice versa):

```
cat file.vcf | java -jar vcf2beagle.jar 0 output_name
java -jar beagle2vcf.jar [.int file] [.markers file] [.bgl file] [.missing file] > output.vcf
```

Using 'Phased' vcf files to find consensus phasing

```
java -jar consensusvcf.jar [vcf 1] [vcf 2] ... > [consensus]
```

Matching SNP positions and genotypes in parent offspring trios (example)

```
perl match.pl -f file2 -g file1 -k 2 -l 2 -v "3 4" > new_file
```

Using BEAGLE for parent-child trio haplotype phasing analysis

```
java -jar beagle.jar trios=<unphased trio file> out=<output_prefix> missing=-
redundant=true iterations=30
```

*beagle.jar, beagle2vcf.jar, consensusvcf.jar are available at the BEAGLE website (at <http://faculty.washington.edu/browning/beagle/beagle.html>)

AutoZplotter.py

```

#Script developed by Dr. Tom Gaunt at the University of Bristol
#!/usr/local/bin/python
import sys
import matplotlib.pyplot as plt

from Tkinter import *
from tkMessageBox import *
from tkColorChooser import askcolor
from tkFileDialog import askopenfilename

def calchet(winhet):
    sum = 0.0
    count = 0.0
    for item in winhet:
        sum += float(item)
        count += 1.0
    return sum/count

#try:
#regfilename = sys.argv[1]
infilename = askopenfilename(defaultextension='.vcf',filetypes=[('Variant Call Format','*.vcf'),('All
files','*.*')])#sys.argv[1]

#reginfile = open(regfilename, 'r')
infile = open(infilename, 'r')
while 1:
    dataline = infile.readline()
    dataline = dataline.strip().split()
    if dataline[0] == "#CHROM":
        break

#results = []
#regions = reginfile.readlines()
#reginfile.close()

hzplotx = []
hzploty = []
hetplotx = []
hetploty = []

plotx = []
ploty = []

currentchr = "null"
counter = 0
windowhet = []

for i in range(25):
    plotx.append([])
    ploty.append([])

while 1:
    dataline = infile.readline()
    #print dataline
    dataline = dataline.strip().split()

```

```

if len(dataline) < 5:
    break
genocodes = dataline[9].split(":")
chrom = dataline[0]
if chrom in
["1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","24","X"
,"Y",1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24]:
    chromn = chrom
else:
    chromn = chrom[3:]
if chromn == "X" or chrom == "X":
    chromn = 23
elif chromn == "Y" or chrom == "Y":
    chromn = 24
else:
    chromn = int(chromn)
if chromn != currentchr:
    currentchr = chromn
    windowhet = []
    counter = 0

pos = dataline[1]
genotype = genocodes[0]
if genotype == "0/0" or genotype == "1/1":
    thishet = 0.0
else:
    thishet = 1.0
if dataline[2][0:2] == "rs":
    windowhet.append(thishet)
if counter > 20 and dataline[2][0:2] == "rs":
    null = windowhet.pop(0)
    plotx[chromn].append(pos)
    thisval = float(calchet(windowhet))
    ploty[chromn].append(float(chromn)*2+thisval)
counter += 1

readdepth = genocodes[1]
#DP = genocodes[2]
#GQ = genocodes[3]
#PL = genocodes[4]
#labelled = 0
#for region in regions:
#    region = region.strip().split()
#    if region[1] == chrom and pos >= region[2] and pos <= region[3]:
#
results.append([chrom,pos,dataline[2]+(" "+dataline[3]+"/ "+dataline[4]+"),genotype,readdep
th,DP,GQ,PL,region[0],region[1],region[2],region[3]])
#    labelled = 1
#if labelled == 0:
#
results.append([chrom,pos,dataline[2]+(" "+dataline[3]+"/ "+dataline[4]+"),genotype,readdep
th,DP,GQ,PL,"NA","NA","NA","NA"])
if (genotype == "0/0" or genotype == "1/1" or genotype == "0|0" or genotype == "1|1") and
dataline[2][0:2] == "rs":

```

```

        hzplotx.append(pos)
        hzploty.append(float(chromn)*2+1+0.4)
    elif dataline[2][0:2] == "rs":
        hetplotx.append(pos)
        hetploty.append(float(chromn)*2+1+0.6)
    #print pos,chrom,genotype

fig = plt.figure()
ax = fig.add_subplot(111)

filelist = infile.name.split("/")
print filelist
filetitle = filelist[-1]
fig.suptitle("File: " + filetitle, fontsize=11)

#for i in range(1,25):
#    ax.plot([0,250000000],[i,i],'k:')

plt.plot(hzplotx,hzploty,'ro',hetplotx,hetploty,'go')
#print "ready",plotx[0][0:5],ploty[0][0:5]

#ax.set_yticks((1.5,2.5,3.5,4.5,5.5,6.5,7.5,8.5,9.5,10.5,11.5,12.5,13.5,14.5,15.5,16.5,17.5,18.5,19.5,20.5,21.5,2
2.5,23.5,24.5))
ytick = []
for i in range(1,25):
    ytick.append(i*2+1)
ax.set_yticks(ytick)
ticklabels = []
for i in range(1,23):
    ticklabels.append("Chr "+str(i))
ticklabels.append("Chr X")
ticklabels.append("Chr Y")

labels = ax.set_yticklabels(ticklabels)

for i in range(len(plotx)):
    ax.plot([0,250000000],[i,i],'k:')
    ax.plot([0,250000000],[i,i],'k-')
    ax.plot(plotx[i],ploty[i],'b-')
print "ready",plotx[0][0:5],ploty[0][0:5]

#ax.axis('tight')
#plt.plot([1,4,8],[2,2,1])
plt.show()
print "done"
infile.close()
#for item in results:
#    print "\t".join(item)

```

```
#except:
#     print "\nscanautoz: A tool to scan autozygous regions from a vcf file\nUsage:\n\tscanautoz\n\t<vcf-file>\n"
```

Missense2fathmm.py

```
#converts missense.vep files to FATHMM input format
WorkingDirectory = '/data/home/user/to/variant_effect_predictor/'
import sys
for id in ["file"]:
    missense_file = open(WorkingDirectory + id + '.vep.missense', 'r')
    fathmm_file = open(WorkingDirectory + id + '_FATHMM.txt', 'w')
    for line in missense_file:
        line = line.replace('\n', '')
        record = line.split('\t')
        aa_pos = record[9]
        snp_col = record[10]
        record_snp = snp_col.split('/')
        snp1 = record_snp[0]
        snp2 = record_snp[1]
        ensp_col = record[13]
        record = ensp_col.split(';')
        ensp = record[0][5:]
        fathmm_file.write(str(ensp) + '\t' + str(snp1)+str(aa_pos)+str(snp2) + '\n')
```


Allele_remover.py

#Initially this script deletes all the heterozygous loci in the proband_file; then the homozygous loci present in the throw_file from the proband_file

```
WorkingDirectory = '/home/path/to/family_data/'
```

```
proband_file = open(WorkingDirectory + 'genome_patient1.txt', 'r')
```

```
throw_file = open(WorkingDirectory + 'patient2_homozygote.txt', 'r')
```

```
output = open(WorkingDirectory + "probandfinal.txt", "w")
```

```
Throw = []
```

```
for record in throw_file:
```

```
    Throw.append(record.strip())
```

```
for record in proband_file:
```

```
    if not record.startswith("#"):
```

```
        record = record.strip().split("\t")
```

```
        if record[3][0] == record[3][1]:
```

```
            if not "\t".join(record) in Throw:
```

```
                output.write("\t".join(record) + "\n")
```

PHI_SO_Terms.txt

```
#Ensembl VEP sequence ontology (SO) terms of predicted high impact (PHI)
mutations (see Alsaadi and Erzurumluoglu et al, 2014 for example) - SO terms
correct as of March 2015
#Can be used when 'greping' out PHI mutations from Ensembl VEP annotated VCF
files (command: grep -f PHI_SO_terms.txt file_name.vep)
splice_acceptor_variant
splice_donor_variant
stop_lost
stop_gained
missense_variant
initiator_codon_variant
inframe_insertion
inframe_deletion
frameshift_variant
transcript_ablation
stop_lost
```

SO_Terms_Indel.txt

```
inframe_insertion
inframe_deletion
frameshift_variant
splice_donor_variant
splice_acceptor_variant
stop_gained
stop_lost
```

Homozygote_alleles.txt

```
AA
GG
CC
TT
-- --
```

Convert_23andme_to_plink.pl

```
#Script made available under Creative Commons Licence by author at:
#https://github.com/thejeshgn/dna/blob/master/code/CONVERT_23AME_PED.pl
#!/usr/bin/perl
# file and ped file format stuff passed in
my($file,$individualID,$familyID,$patID,$matID,$sex,$phen)=@ARGV;

# make sure vars are passed correctly
if(scalar(@ARGV) == 0){
    print "\nParameters need to be passed for the script to work: CONVERT_23AME_PED.pl
genome.txt MyID MyFamID\n";
    print "Please pass parameters into the script of the following form:\n\n";
    print "File name: required\n";
    print "Individual ID: optional, if not provided will be set to ID001\n";
    print "Family ID: optional, if not provided will be set to FAM001\n";
    print "Paternal ID: if not provided will be set to unknown\n";
    print "Maternal ID: if not provided will be set to unknown\n";
    print "Sex: 1 = male, 2 = female, if not provided will be set to unknown\n";
    print "Phenotype: 1, or 2, if not set will be set to -9\n\n";
}#END IF
else{
    # set default bars if needed
    if(!$individualID){
        $individualID = "ID001";
    }
    if(!$familyID){
        $familyID = "FAM001";
    }
    if(!$patID){
        $patID = "unknown";
    }
    if(!$matID){
        $matID = "unknown";
    }
    if(!$sex){
        $sex = "unknown";
    }
    if(!$phen){
        $phen = "-9";
    }
    # make sure file exists
    if (-e $file)
    {
        # stuff before the genotypes
        my $initialSegment = "$familyID\t$individualID\t$patID\t$matID\t$sex\t$phen";
        # now add genotype info
        makePED($file,$initialSegment)
    }
    else
    {
        print "\nPlease enter a file name which exists\n\n";
    }
}# do stuff
sub makePED{
```

```

my $rawfilein = shift;
my $presection = shift;
open(RAWDATA, "<", "$rawfilein") or die $!;
# split
my @splitfile = split(/\./,$rawfilein);
my $rawfile = $splitfile[0];
open(MAP, ">", $rawfile.".map") or die $!;
open (PED, ">", $rawfile.".ped") or die $!;
my @line;
my @genotype;
print PED $presection;
while(my $this_line = <RAWDATA>) {
    if(($this_line =~ /^rs/) || ($this_line =~ /^i/)){
        @line = split('\t', $this_line);
        @genotype = unpack('aa', $line[3]);
        if ($genotype[0] eq '-' or $genotype[0] eq 'T' or $genotype[0] eq 'D')
        {
            @genotype = unpack('aa', "00");
        }
        print MAP "$line[1] $line[0] 0 $line[2]\n";
        if ($line[1] eq 'X' or $line[1] eq 'MT' or $line[1] eq 'Y') {
            print PED " $genotype[0] 0"; } else { print PED "t$genotype[0] $genotype[1]"; }
        }
    }
}
close PED;
close MAP;
close RAWDATA;
}

```

Convert_plink_to_vcf.py

```
#Script made available under Creative Commons Licence by author at:
#https://github.com/chapmanb/bcbio-nextgen/blob/master/scripts/utils/plink_to_vcf.py
#!/usr/bin/env python
"""Convert Plink ped/map files into VCF format using plink and Plink/SEQ.
Requires:
plink: http://pngu.mgh.harvard.edu/~purcell/plink/
PLINK/SEQ: http://atgu.mgh.harvard.edu/plinkseq/
bx-python: https://bitbucket.org/james_taylor/bx-python/wiki/Home
You also need the genome reference file in 2bit format:
http://genome.ucsc.edu/FAQ/FAQformat.html#format7
Usage:
    plink_to_vcf.py <ped file> <map file> <UCSC reference file in 2bit format>
"""
import os
import sys
import subprocess
from bx.seq import twobit
def main(ped_file, map_file, ref_file):
    base_dir = os.getcwd()
    pbed_prefix = convert_to_plink_bed(ped_file, map_file, base_dir)
    vcf_file = convert_bed_to_vcf(pbed_prefix, ped_file, base_dir)
    fix_nonref_positions(vcf_file, ref_file)
def convert_to_plink_bed(ped_file, map_file, base_dir):
    # from ubuntu package, 'plink' otherwise
    for plink_cl in ["p-link", "plink"]:
        try:
            subprocess.check_call([plink_cl, "--help"])
            break
        except:
            pass
    plink_prefix = os.path.splitext(os.path.basename(ped_file))[0]
    work_dir = os.path.join(base_dir, "vcfconvert")
    if not os.path.exists(work_dir):
        os.makedirs(work_dir)
    out_base = os.path.join(work_dir, plink_prefix)
    if not os.path.exists("{0}.bed".format(out_base)):
        subprocess.check_call([plink_cl, "--noweb",
                               "--ped", ped_file, "--map", map_file,
                               "--make-bed", "--out", out_base])
    return out_base
def convert_bed_to_vcf(pbed_prefix, ped_file, base_dir):
    out_file = os.path.join(base_dir,
                            "{0}.vcf".format(os.path.splitext(os.path.basename(ped_file))[0]))
    if not os.path.exists(out_file):
        subprocess.check_call(["./pseq", pbed_prefix, "new-project"])
        subprocess.check_call(["./pseq", pbed_prefix, "load-plink",
                               "--file", pbed_prefix, "--id", "vcfconvert"])
        with open(out_file, "w") as out_handle:
            subprocess.check_call(["./pseq", pbed_prefix, "write-vcf"],
                                stdout=out_handle)
    return out_file
def fix_line_problems(parts):
    """Fix problem alleles and reference/variant bases in VCF line.
```

```

"""
varinfo = parts[:9]
genotypes = []
# replace haploid calls
for x in parts[9:]:
    if len(x) == 1:
        x = "./."
    genotypes.append(x)
if varinfo[3] == "0": varinfo[3] = "N"
if varinfo[4] == "0": varinfo[4] = "N"
return varinfo, genotypes
def fix_vcf_line(parts, ref_base):
    """Orient VCF allele calls with respect to reference base.
    Handles cases with ref and variant swaps. strand complements.
    """
    swap = {"1/1": "0/0", "0/1": "0/1", "0/0": "1/1", "./.": "./."}
    complements = {"G": "C", "A": "T", "C": "G", "T": "A", "N": "N"}
    varinfo, genotypes = fix_line_problems(parts)
    ref, var = varinfo[3:5]
    # non-reference regions or non-informative, can't do anything
    if ref_base in [None, "N"] or set(genotypes) == set(["./."]):
        varinfo = None
    # matching reference, all good
    elif ref_base == ref:
        assert ref_base == ref, (ref_base, parts)
    # swapped reference and alternate regions
    elif ref_base == var or ref in ["N", "0"]:
        varinfo[3] = var
        varinfo[4] = ref
        genotypes = [swap[x] for x in genotypes]
    # reference is on alternate strand
    elif ref_base != ref and complements[ref] == ref_base:
        varinfo[3] = complements[ref]
        varinfo[4] = complements[var]
    # unspecified alternative base
    elif ref_base != ref and var in ["N", "0"]:
        varinfo[3] = ref_base
        varinfo[4] = ref
        genotypes = [swap[x] for x in genotypes]
    # swapped and on alternate strand
    elif ref_base != ref and complements[var] == ref_base:
        varinfo[3] = complements[var]
        varinfo[4] = complements[ref]
        genotypes = [swap[x] for x in genotypes]
    else:
        print "Did not associate ref {0} with line: {1}".format(
            ref_base, varinfo)
    if varinfo is not None:
        return varinfo + genotypes
def fix_nonref_positions(in_file, ref_file):
    """Fix Genotyping VCF positions where the bases are all variants.
    The plink/pseq output does not handle these correctly, and
    has all reference/variant bases reversed.
    """
    ignore_chrs = ["."]

```

```

ref2bit = twobit.TwoBitFile(open(ref_file))
out_file = apply("{0}-fix{1}".format, os.path.splitext(in_file))

with open(in_file) as in_handle:
    with open(out_file, "w") as out_handle:
        for line in in_handle:
            if line.startswith("#"):
                out_handle.write(line)
            else:
                parts = line.rstrip("\r\n").split("\t")
                pos = int(parts[1])
                # handle chr/non-chr naming
                if parts[0] not in ref2bit.keys():
                    parts[0] = parts[0].replace("chr", "")
                ref_base = None
                if parts[0] not in ignore_chrs:
                    try:
                        ref_base = ref2bit[parts[0]].get(pos-1, pos).upper()
                    except Exception, msg:
                        # off the end of the chromosome
                        if str(msg).startswith("end before start"):
                            print msg
                        else:
                            print parts
                            raise
                parts = fix_vcf_line(parts, ref_base)
                if parts is not None:
                    out_handle.write("\t".join(parts) + "\n")
        return out_file
if __name__ == "__main__":
    if len(sys.argv) != 4:
        print "Incorrect arguments"
        print __doc__
        sys.exit(1)
    main(*sys.argv[1:])

```

BIBLIOGRAPHY

1. Savage DC. Microbial ecology of the gastrointestinal tract. Annual review of microbiology. 1977;31:107-33. Epub 1977/01/01. doi: 10.1146/annurev.mi.31.100177.000543. PubMed PMID: 334036.
2. Guertin DA, Sabatini DM. Cell Size Control. 2006. doi: 10.1038/npg.els.0003359.
3. Wallace DC. A mitochondrial paradigm for degenerative diseases and ageing. Novartis Foundation symposium. 2001;235:247-63; discussion 63-6. Epub 2001/03/31. PubMed PMID: 11280029.
4. Reiling E, van Vliet-Ostaptchouk JV, van 't Riet E, van Haeften TW, Arp PA, Hansen T, et al. Genetic association analysis of 13 nuclear-encoded mitochondrial candidate genes with type II diabetes mellitus: the DAMAGE study. European journal of human genetics : EJHG. 2009;17(8):1056-62. Epub 2009/02/12. doi: 10.1038/ejhg.2009.4. PubMed PMID: 19209188; PubMed Central PMCID: PMC2986549.
5. Chinnery PF. Searching for nuclear-mitochondrial genes. Trends in genetics : TIG. 2003;19(2):60-2. Epub 2003/01/28. PubMed PMID: 12547509.
6. Dictionaries O. Nucleus: Oxford University Press; 2013 [October 15, 2013]. Available from: <http://oxforddictionaries.com/definition/english/nucleus>
7. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171(4356):737-8.
8. Levene PA. THE STRUCTURE OF YEAST NUCLEIC ACID: IV. AMMONIA HYDROLYSIS. Journal of Biological Chemistry. 1919;40(2):415-24.
9. Tjio JH, Puck TT. Genetics of somatic mammalian cells. II. Chromosomal constitution of cells in tissue culture. The Journal of experimental medicine. 1958;108(2):259-68. Epub 1958/08/01. PubMed PMID: 13563760; PubMed Central PMCID: PMC2136870.
10. Mazat JP, Ransac S, Heiske M, Devin A, Rigoulet M. Mitochondrial energetic metabolism-some general principles. IUBMB life. 2013;65(3):171-9. Epub 2013/02/27. doi: 10.1002/iub.1138. PubMed PMID: 23441039.
11. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 2013;494(7435):77-80. doi: <http://www.nature.com/nature/journal/v494/n7435/abs/nature11875.html#supplementary-information>.
12. Mayr E. The Growth of Biological Thought: Diversity, Evolution, and Inheritance: Belknap Press of Harvard University Press; 1982.
13. Zirkle C. Natural Selection before the "Origin of Species". Proceedings of the American Philosophical Society. 1941;84(1):71-123. doi: 10.2307/984852.
14. Cosman MP, Jones LG. Handbook to Life in the Medieval World, 3-Volume Set: Facts On File, Incorporated; 2009.
15. Mendel G. Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver Brunn. 1866;4:3-47.

16. Miescher F. Ueber die chemische Zusammensetzung der Eiterzellen. Monographie. 1872.
17. Miko I. Thomas Hunt Morgan and sex linkage. *Scitable*. 2008;(1):1.
18. Avery OT, Macleod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of experimental medicine*. 1944;79(2):137-58. Epub 1944/02/01. PubMed PMID: 19871359; PubMed Central PMCID: PMCPMC2135445.
19. Crick F. Central dogma of molecular biology. *Nature*. 1970;227(5258):561-3. Epub 1970/08/08. PubMed PMID: 4913914.
20. Niu DK, Jiang L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and biophysical research communications*. 2013;430(4):1340-3. Epub 2012/12/27. doi: 10.1016/j.bbrc.2012.12.074. PubMed PMID: 23268340.
21. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol*. 2012;30(11):1095-106. Epub 2012/11/10. doi: 10.1038/nbt.2422. PubMed PMID: 23138309; PubMed Central PMCID: PMCPMC3703467.
22. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. *Nature reviews Genetics*. 2010;11(8):559-71. Epub 2010/07/16. doi: 10.1038/nrg2814. PubMed PMID: 20628352.
23. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(49):19428-33. Epub 2007/11/28. doi: 10.1073/pnas.0709013104. PubMed PMID: 18040051; PubMed Central PMCID: PMC2148306.
24. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. *Genome research*. 2007;17(6):669-81. Epub 2007/06/15. doi: 10.1101/gr.6339607. PubMed PMID: 17567988.
25. Thomas DJ, Rosenbloom KR, Clawson H, Hinrichs AS, Trumbower H, Raney BJ, et al. The ENCODE Project at UC Santa Cruz. *Nucleic acids research*. 2007;35(Database issue):D663-7. Epub 2006/12/15. doi: 10.1093/nar/gkl1017. PubMed PMID: 17166863; PubMed Central PMCID: PMC1781110.
26. Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306(5696):636-40. doi: 10.1126/science.1105136.
27. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-8. Epub 2010/09/03. doi: 10.1038/nature09298. PubMed PMID: 20811451; PubMed Central PMCID: PMC3173859.
28. Consortium TIH. The International HapMap Project. *Nature*. 2003;426(6968):789-96. doi: http://www.nature.com/nature/journal/v426/n6968/supinfo/nature02168_S1.html.

29. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*. 2010;328(5978):636-9. Epub 2010/03/12. doi: 10.1126/science.1186802. PubMed PMID: 20220176; PubMed Central PMCID: PMC23037280.
30. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics*. 2011;43(7):712-4. Epub 2011/06/15. doi: 10.1038/ng.862. PubMed PMID: 21666693; PubMed Central PMCID: PMC3322360.
31. Duggal NK, Emerman M. Evolutionary conflicts between viruses and restriction factors shape immunity. *Nature reviews Immunology*. 2012;12(10):687-95. Epub 2012/09/15. doi: 10.1038/nri3295. PubMed PMID: 22976433; PubMed Central PMCID: PMC3690816.
32. Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, et al. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science*. 1988;242(4884):1427-30. Epub 1988/12/09. PubMed PMID: 3201231.
33. Guichard C, Harricane MC, Lafitte JJ, Godard P, Zaegel M, Tack V, et al. Axonemal dynein intermediate-chain gene (DNAI1) mutations result in situs inversus and primary ciliary dyskinesia (Kartagener syndrome). *American journal of human genetics*. 2001;68(4):1030-5. Epub 2001/03/07. doi: 10.1086/319511. PubMed PMID: 11231901; PubMed Central PMCID: PMC1275621.
34. Mitchison HM, Schmidts M, Loges NT, Freshour J, Dritsoula A, Hirst RA, et al. Mutations in axonemal dynein assembly factor DNAAF3 cause primary ciliary dyskinesia. *Nature genetics*. 2012;44(4):381-9, S1-2. Epub 2012/03/06. doi: 10.1038/ng.1106. PubMed PMID: 22387996; PubMed Central PMCID: PMC3315610.
35. Kott E, Legendre M, Copin B, Papon JF, Dastot-Le Moal F, Montantin G, et al. Loss-of-Function Mutations in RSPH1 Cause Primary Ciliary Dyskinesia with Central-Complex and Radial-Spoke Defects. *American journal of human genetics*. 2013;93(3):561-70. Epub 2013/09/03. doi: 10.1016/j.ajhg.2013.07.013. PubMed PMID: 23993197.
36. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.
37. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, et al. A physical map of the human genome. *Nature*. 2001;409(6822):934-41. Epub 2001/03/10. doi: 10.1038/35057157. PubMed PMID: 11237014.
38. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409(6822):928-33. Epub 2001/03/10. doi: 10.1038/35057149. PubMed PMID: 11237013.
39. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304-51. Epub 2001/02/22. doi: 10.1126/science.1058040. PubMed PMID: 11181995.
40. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? *J Med Genet*. 2011;48(9):580-9. Epub 2011/07/07. doi: 10.1136/jmedgenet-2011-100223. PubMed PMID: 21730106.

41. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotechnol.* 2012;30(4):295-6. Epub 2012/04/12. doi: 10.1038/nbt0412-295. PubMed PMID: 22491260.
42. Khorana HG, Buchi H, Ghosh H, Gupta N, Jacob TM, Kossel H, et al. Polynucleotide synthesis and the genetic code. *Cold Spring Harbor symposia on quantitative biology.* 1966;31:39-49. Epub 1966/01/01. PubMed PMID: 5237635.
43. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F, et al. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America.* 1965;53(5):1161-8. Epub 1965/05/01. PubMed PMID: 5330357; PubMed Central PMCID: PMC301388.
44. Tsongalis GJ, Silverman LM. Molecular pathology of the fragile X syndrome. *Archives of pathology & laboratory medicine.* 1993;117(11):1121-5. Epub 1993/11/01. PubMed PMID: 8239933.
45. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science.* 2013;341(6141):1237758. Epub 2013/07/06. doi: 10.1126/science.1237758. PubMed PMID: 23828942.
46. Driscoll DA, Gross S. Prenatal Screening for Aneuploidy. *New England Journal of Medicine.* 2009;360(24):2556-62. doi: doi:10.1056/NEJMcp0900134. PubMed PMID: 19516035.
47. Unit IMIMMPPHCMCEC, Medicine dBooP, Public Health UA, Hill PEUNCC. *A Dictionary of Epidemiology: Oxford University Press, USA; 2008.*
48. Lewis SJ, Zuccolo L, Davey Smith G, Macleod J, Rodriguez S, Draper ES, et al. Fetal alcohol exposure and IQ at age 8: evidence from a population-based birth-cohort study. *PloS one.* 2012;7(11):e49407. Epub 2012/11/21. doi: 10.1371/journal.pone.0049407. PubMed PMID: 23166662; PubMed Central PMCID: PMC3498109.
49. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. *Nature.* 2009;461(7265):724-6. Epub 2009/10/09. doi: 10.1038/461724a. PubMed PMID: 19812653.
50. Heshka JT, Palleschi C, Howley H, Wilson B, Wells PS. A systematic review of perceived risks, psychological and behavioral impacts of genetic testing. *Genet Med.* 2008;10(1):19-32. Epub 2008/01/17. doi: 10.1097/GIM.0b013e31815f524f. PubMed PMID: 18197053.
51. Strachan T, Read AP. *Human Molecular Genetics 4: Garland Science/Taylor & Francis Group; 2011.*
52. Online Mendelian Inheritance in Man OM-NIoGM, Johns Hopkins University (Baltimore, MD), {15/10/12}. World Wide Web URL: <http://omim.org/>. OMIM 2013. Available from: <http://www.omim.org/>.
53. Education CfG. Genetic Conditions - An Overview 2013 [09/10/2013]. Available from: <http://www.genetics.edu.au/Publications-and-Resources/Genetics-Fact-Sheets/Genetic-Conditions-Overview-FS2>.
54. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human mutation.* 2008;29(1):6-13. Epub 2007/11/15. doi: 10.1002/humu.20654. PubMed PMID: 18000842.

55. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. *Nucleic acids research*. 2013;41(D1):D48-D55.
56. Huang L, Wilkinson MF. Regulation of nonsense-mediated mRNA decay. *Wiley interdisciplinary reviews RNA*. 2012;3(6):807-28. Epub 2012/10/03. doi: 10.1002/wrna.1137. PubMed PMID: 23027648.
57. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome research*. 2009;19(9):1553-61. Epub 2009/07/16. doi: 10.1101/gr.092619.109. PubMed PMID: 19602639; PubMed Central PMCID: PMC2752137.
58. Simmons MJ, Crow JF. Mutations affecting fitness in *Drosophila* populations. *Annual review of genetics*. 1977;11:49-78. Epub 1977/01/01. doi: 10.1146/annurev.ge.11.120177.000405. PubMed PMID: 413473.
59. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *European journal of human genetics : EJHG*. 2012;20(5):490-7.
60. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010;141(2):210-7. Epub 2010/04/21. doi: 10.1016/j.cell.2010.03.032. PubMed PMID: 20403315.
61. Simunovic MP. Colour vision deficiency. *Eye (London, England)*. 2010;24(5):747-55. Epub 2009/11/21. doi: 10.1038/eye.2009.251. PubMed PMID: 19927164.
62. Walling HW, Baldassare JJ, Westfall TC. Molecular aspects of Huntington's disease. *Journal of neuroscience research*. 1998;54(3):301-8. Epub 1998/11/18. PubMed PMID: 9819135.
63. Ghumman S, Goel N, Rajaram S, Singh KC, Kansal B, Dewan P. Pregnancy in an achondroplastic dwarf: a case report. *Journal of the Indian Medical Association*. 2005;103(10):536, 8. Epub 2006/02/28. PubMed PMID: 16498757.
64. Berlin AL, Paller AS, Chan LS. Incontinentia pigmenti: a review and update on the molecular basis of pathophysiology. *Journal of the American Academy of Dermatology*. 2002;47(2):169-87; quiz 88-90. Epub 2002/07/26. PubMed PMID: 12140463.
65. Kleopa KA, Scherer SS. Molecular genetics of X-linked Charcot-Marie-Tooth disease. *Neuromolecular medicine*. 2006;8(1-2):107-22. Epub 2006/06/16. doi: 10.1385/nmm:8:1:107. PubMed PMID: 16775370.
66. Carpenter TO, Imel EA, Holm IA, Jan de Beur SM, Insogna KL. A clinician's guide to X-linked hypophosphatemia. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*. 2011;26(7):1381-8. Epub 2011/05/04. doi: 10.1002/jbmr.340. PubMed PMID: 21538511; PubMed Central PMCID: PMC3157040.
67. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*. 2013;339(6121):819-23. Epub 2013/01/05. doi: 10.1126/science.1231143. PubMed PMID: 23287718; PubMed Central PMCID: PMC3795411.
68. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*. 1987;236(4808):1567-70. Epub 1987/06/19. PubMed PMID: 2884728.

69. NHS. Diabetes, type 2: NHS Choices; 2013 [updated 24/07/2012; cited 2013 21-10-13]. Available from: <http://www.nhs.uk/Conditions/Diabetes-type2/Pages/Introduction.aspx>.
70. Gov.uk. Reducing obesity and improving diet: Department of Health; 2013 [updated 25 March 2013; cited 2013 25-10-13]. Available from: <https://www.gov.uk/government/policies/reducing-obesity-and-improving-diet>.
71. Diabetes.co.uk. NHS and Diabetes 2013 [cited 2013 25-10-13]. Available from: <http://www.diabetes.co.uk/nhs/>.
72. Lobo I. Multifactorial Inheritance and Genetic Disease. *Scitable*. 2008;1(1).
73. Bulmer M. Galton's law of ancestral heredity. *Heredity (Edinb)*. 1998;81 (Pt 5):579-85. Epub 1999/02/13. PubMed PMID: 9988590.
74. Pomerantz MM, Freedman ML. The genetics of cancer risk. *Cancer journal (Sudbury, Mass)*. 2011;17(6):416-22. Epub 2011/12/14. doi: 10.1097/PPO.0b013e31823e5387. PubMed PMID: 22157285.
75. Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. *Nature reviews Genetics*. 2009;10(12):872-8. Epub 2009/10/28. doi: 10.1038/nrg2670. PubMed PMID: 19859063.
76. Fisher RA. *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*: Royal Society of Edinburgh; 1918.
77. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53. Epub 2009/10/09. doi: 10.1038/nature08494. PubMed PMID: 19812666; PubMed Central PMCID: PMC2831613.
78. Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106:9362 - 7. PubMed PMID: doi:10.1073/pnas.0903103106.
79. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews Genetics*. 2011;12(11):745-55. Epub 2011/09/29. doi: 10.1038/nrg3031. PubMed PMID: 21946919.
80. Alsaadi MM, Erzurumluoglu AM, Rodriguez S, Guthrie PA, Gaunt TR, Omar HZ, et al. Nonsense mutation in coiled-coil domain containing 151 gene (CCDC151) causes primary ciliary dyskinesia. *Human mutation*. 2014;35(12):1446-8. Epub 2014/09/17. doi: 10.1002/humu.22698. PubMed PMID: 25224326.
81. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(23):9362-7. Epub 2009/05/29. doi: 10.1073/pnas.0903103106. PubMed PMID: 19474294; PubMed Central PMCID: PMC2687147.
82. Stitzel NO, Kiezun A, Sunyaev SR. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol*. 2011;12:227.
83. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS biology*. 2010;8(1):e1000294. Epub

- 2010/02/04. doi: 10.1371/journal.pbio.1000294. PubMed PMID: 20126254; PubMed Central PMCID: PMC2811148.
84. Gibson G. Rare and common variants: twenty arguments. *Nature reviews Genetics*. 2011;13(2):135-45. Epub 2012/01/19. doi: 10.1038/nrg3118. PubMed PMID: 22251874.
 85. Mossey PA. The heritability of malocclusion: Part 1--Genetics, principles and terminology. *British journal of orthodontics*. 1999;26(2):103-13. Epub 1999/07/27. PubMed PMID: 10420244.
 86. Bittles AH, Black ML. Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*. 2010;107(suppl 1):1779-86. doi: 10.1073/pnas.0906079106.
 87. Day IN. dbSNP in the detail and copy number complexities. *Human mutation*. 2010;31(1):2-4. Epub 2009/12/22. doi: 10.1002/humu.21149. PubMed PMID: 20024941.
 88. Alharbi KK, Aldahmesh MA, Spanakis E, Haddad L, Whittall RA, Chen XH, et al. Mutation scanning by meltMADGE: validations using BRCA1 and LDLR, and demonstration of the potential to identify severe, moderate, silent, rare, and paucimorphic mutations in the general population. *Genome research*. 2005;15(7):967-77. Epub 2005/07/07. doi: 10.1101/gr.3313405. PubMed PMID: 15998910; PubMed Central PMCID: PMC1172041.
 89. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics*. 2008;9(5):356-69. Epub 2008/04/10. doi: 10.1038/nrg2344. PubMed PMID: 18398418.
 90. Lander ES. The new genomics: global views of biology. *Science*. 1996;274(5287):536-9. Epub 1996/10/25. PubMed PMID: 8928008.
 91. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends in genetics : TIG*. 2001;17(9):502-10. Epub 2001/08/30. PubMed PMID: 11525833.
 92. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*. 2002;11:2417-23.
 93. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *American journal of human genetics*. 2000;67(4):881-900. Epub 2000/09/14. doi: 10.1086/303070. PubMed PMID: 10986041; PubMed Central PMCID: PMC1287893.
 94. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69:124-37.
 95. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*. 2009;19(3):212-9. Epub 2009/06/02. doi: 10.1016/j.gde.2009.04.010. PubMed PMID: 19481926; PubMed Central PMCID: PMC2914559.
 96. Aidoo M, Terlouw DJ, Kolczak MS, McElroy PD, ter Kuile FO, Kariuki S, et al. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet*. 2002;359(9314):1311-2. Epub 2002/04/20. doi: 10.1016/s0140-6736(02)08273-9. PubMed PMID: 11965279.

97. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823-8. Epub 2012/02/22. doi: 10.1126/science.1215040. PubMed PMID: 22344438; PubMed Central PMCID: PMC3299548.
98. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nature reviews Genetics*. 2012;13(8):565-75. Epub 2012/07/19. doi: 10.1038/nrg3241. PubMed PMID: 22805709.
99. Evans DM, Purcell S. Power calculations in genetic studies. *Cold Spring Harbor protocols*. 2012;2012(6):664-74. Epub 2012/06/05. doi: 10.1101/pdb.top069559. PubMed PMID: 22661434.
100. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *American journal of human genetics*. 2010;86(6):929-42. Epub 2010/06/22. doi: 10.1016/j.ajhg.2010.05.002. PubMed PMID: 20560208; PubMed Central PMCID: PMC3032061.
101. Neale BM. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011;7:e1001322.
102. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*. 2010;42(7):565-9. doi: http://www.nature.com/ng/journal/v42/n7/supinfo/ng.608_S1.html.
103. Tranah GJ. Mitochondrial-nuclear epistasis: implications for human aging and longevity. *Ageing research reviews*. 2011;10(2):238-52. Epub 2010/07/06. doi: 10.1016/j.arr.2010.06.003. PubMed PMID: 20601194; PubMed Central PMCID: PMC2995012.
104. Online EB. Consanguinity 2013 [cited 2013 16 September]. Available from: <http://www.britannica.com/EBchecked/topic/133242/consanguinity>.
105. Hamamy H, Antonarakis SE, Cavalli-Sforza LL, Temtamy S, Romeo G, Kate LP, et al. Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report. *Genet Med*. 2011;13(9):841-7. Epub 2011/05/11. doi: 10.1097/GIM.0b013e318217477f. PubMed PMID: 21555946.
106. Ceballos FC, Alvarez G. Royal dynasties as human inbreeding laboratories: the Habsburgs. *Heredity (Edinb)*. 2013;111(2):114-21. Epub 2013/04/11. doi: 10.1038/hdy.2013.25. PubMed PMID: 23572123; PubMed Central PMCID: PMC3716267.
107. Hussain R, Bittles AH. Consanguineous marriage and differentials in age at marriage, contraceptive use and fertility in Pakistan. *Journal of biosocial science*. 1999;31(1):121-38. Epub 1999/03/19. PubMed PMID: 10081242.
108. BITTLES AH. WHEN COUSINS MARRY: A REVIEW OF CONSANGUINITY IN THE MIDDLE EAST. *Perspectives in Human Biology* 1995. p. 71-83.
109. Ortiz F. Cuban counterpoint, tobacco and sugar: Duke University Press; 1995.
110. Khlat M. Consanguineous marriage and reproduction in Beirut, Lebanon. *American journal of human genetics*. 1988;43(2):188-96. Epub 1988/08/01. PubMed PMID: 3400644; PubMed Central PMCID: PMC3715345.
111. Vardi-Saliternik R, Friedlander Y, Cohen T. Consanguinity in a population sample of Israeli Muslim Arabs, Christian Arabs and Druze. *Annals of human*

- biology. 2002;29(4):422-31. Epub 2002/08/06. doi: 10.1080/03014460110100928. PubMed PMID: 12160475.
112. Khoury SA, Massad D. Consanguineous marriage in Jordan. *American journal of medical genetics*. 1992;43(5):769-75. Epub 1992/07/15. doi: 10.1002/ajmg.1320430502. PubMed PMID: 1642259.
113. Tadmouri GO, Nair P, Obeid T, Al Ali MT, Al Khaja N, Hamamy HA. Consanguinity and reproductive health among Arabs. *Reproductive health*. 2009;6:17. Epub 2009/10/09. doi: 10.1186/1742-4755-6-17. PubMed PMID: 19811666; PubMed Central PMCID: PMC2765422.
114. Zlotogora J, Hujerat Y, Zalman L, Barges S, Filon D, Koren A, et al. Origin and expansion of four different beta globin mutations in a single Arab village. *American journal of human biology : the official journal of the Human Biology Council*. 2005;17(5):659-61. Epub 2005/09/02. doi: 10.1002/ajhb.20429. PubMed PMID: 16136542.
115. Miller EN, Fadl M, Mohamed HS, Elzein A, Jamieson SE, Cordell HJ, et al. Y chromosome lineage- and village-specific genes on chromosomes 1p22 and 6q27 control visceral leishmaniasis in Sudan. *PLoS genetics*. 2007;3(5):e71. Epub 2007/05/16. doi: 10.1371/journal.pgen.0030071. PubMed PMID: 17500593; PubMed Central PMCID: PMCPMC1866354.
116. Aldahmesh MA, Abu-Safieh L, Khan AO, Al-Hassnan ZN, Shaheen R, Rajab M, et al. Allelic heterogeneity in inbred populations: the Saudi experience with Alstrom syndrome as an illustrative example. *American journal of medical genetics Part A*. 2009;149A(4):662-5. Epub 2009/03/14. doi: 10.1002/ajmg.a.32753. PubMed PMID: 19283855.
117. Zlotogora J. The molecular basis of autosomal recessive diseases among the Arabs and Druze in Israel. *Human genetics*. 2010;128(5):473-9. Epub 2010/09/21. doi: 10.1007/s00439-010-0890-8. PubMed PMID: 20852892.
118. Teebi AS, Teebi SA. Genetic diversity among the Arabs. *Community genetics*. 2005;8(1):21-6. Epub 2005/03/16. doi: 10.1159/000083333. PubMed PMID: 15767750.
119. Zlotogora J, Hujerat Y, Barges S, Shalev SA, Chakravarti A. The fate of 12 recessive mutations in a single village. *Annals of human genetics*. 2007;71(Pt 2):202-8. Epub 2007/03/03. doi: 10.1111/j.1469-1809.2006.00308.x. PubMed PMID: 17331080.
120. Charlesworth D, Willis JH. The genetics of inbreeding depression. *Nature reviews Genetics*. 2009;10(11):783-96. Epub 2009/10/17. doi: 10.1038/nrg2664. PubMed PMID: 19834483.
121. Biemont C. Inbreeding effects in the epigenetic era. *Nature reviews Genetics*. 2010;11(3):234. Epub 2010/01/29. doi: 10.1038/nrg2664-c1. PubMed PMID: 20107433.
122. Zhang HY, He H, Chen LB, Li L, Liang MZ, Wang XF, et al. A genome-wide transcription analysis reveals a close correlation of promoter INDEL polymorphism and heterotic gene expression in rice hybrids. *Molecular plant*. 2008;1(5):720-31. Epub 2009/10/15. doi: 10.1093/mp/ssn022. PubMed PMID: 19825576.
123. Postma E, Martini L, Martini P. Inbred women in a small and isolated Swiss village have fewer children. *Journal of evolutionary biology*. 2010;23(7):1468-74. Epub 2010/05/25. doi: 10.1111/j.1420-9101.2010.02013.x. PubMed PMID: 20492085.

124. Nebert DW, Galvez-Peralta M, Shi Z, Dragin N. Inbreeding and epigenetics: beneficial as well as deleterious effects. *Nature reviews Genetics*. 2010;11(9):662. Epub 2010/07/28. doi: 10.1038/nrg2664-c2. PubMed PMID: 20661256; PubMed Central PMCID: PMCPMC3025405.
125. Darr A, Modell B. The frequency of consanguineous marriage among British Pakistanis. *J Med Genet*. 1988;25(3):186-90. Epub 1988/03/01. PubMed PMID: 3351906; PubMed Central PMCID: PMCPMC1015484.
126. Bittles AH, Neel JV. The costs of human inbreeding and their implications for variations at the DNA level. *Nature genetics*. 1994;8(2):117-21. Epub 1994/10/01. doi: 10.1038/ng1094-117. PubMed PMID: 7842008.
127. Hoodfar E, Teebi AS. Genetic referrals of Middle Eastern origin in a western city: inbreeding and disease profile. *J Med Genet*. 1996;33(3):212-5. Epub 1996/03/01. PubMed PMID: 8728693; PubMed Central PMCID: PMCPMC1051869.
128. Ober C, Elias S, Kostyu DD, Hauck WW. Decreased fecundability in Hutterite couples sharing HLA-DR. *American journal of human genetics*. 1992;50(1):6-14. Epub 1992/01/01. PubMed PMID: 1729895; PubMed Central PMCID: PMCPMC1682532.
129. Gnanalingham MG, Gnanalingham KK, Singh A. Congenital heart disease and parental consanguinity in South India. *Acta paediatrica (Oslo, Norway : 1992)*. 1999;88(4):473-4. Epub 1999/05/26. PubMed PMID: 10342554.
130. Grant JC, Bittles AH. The comparative role of consanguinity in infant and childhood mortality in Pakistan. *Annals of human genetics*. 1997;61(Pt 2):143-9. Epub 1997/03/01. doi: 10.1046/j.1469-1809.1997.6120143.x. PubMed PMID: 9177121.
131. Stromme P, Suren P, Kanavin OJ, Rootwelt T, Woldseth B, Abdelnoor M, et al. Parental consanguinity is associated with a seven-fold increased risk of progressive encephalopathy: a cohort study from Oslo, Norway. *European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society*. 2010;14(2):138-45. Epub 2009/05/19. doi: 10.1016/j.ejpn.2009.03.007. PubMed PMID: 19446480.
132. Ober C, Hyslop T, Hauck WW. Inbreeding effects on fertility in humans: evidence for reproductive compensation. *American journal of human genetics*. 1999;64(1):225-31. Epub 1999/01/23. doi: 10.1086/302198. PubMed PMID: 9915962; PubMed Central PMCID: PMCPMC1377721.
133. Bittles AH. A Background Summary of Consanguineous Marriages. Available online: <http://consangnet/images/d/dd/01AHBWeb3pdf>. 2001.
134. Fuster V. Inbreeding pattern and reproductive success in a rural community from Galicia (Spain). *Journal of biosocial science*. 2003;35(1):83-93. Epub 2003/01/23. PubMed PMID: 12537158.
135. Helgason A, Palsson S, Gudbjartsson DF, Kristjansson T, Stefansson K. An association between the kinship and fertility of human couples. *Science*. 2008;319(5864):813-6. Epub 2008/02/09. doi: 10.1126/science.1150232. PubMed PMID: 18258915.
136. Bittles AH, Black ML. The impact of consanguinity on neonatal and infant health. *Early human development*. 2010;86(11):737-41. Epub 2010/09/14. doi: 10.1016/j.earlhumdev.2010.08.003. PubMed PMID: 20832202.

137. Tuncbilek E, Koc I. Consanguineous marriage in Turkey and its impact on fertility and mortality. *Annals of human genetics*. 1994;58(Pt 4):321-9. Epub 1994/10/01. PubMed PMID: 7864588.
138. Bittles AH, Savithri HS, Venkatesha Murthy H, Wang W, Cahill J, Baskaran G, et al. Consanguineous marriage, a familiar story full of surprises. In: Macbeth H, Shetty P, editors. *Health and Ethnicity*: Taylor & Francis; 2001. p. 68-78.
139. Bunday S, Alam H. A five-year prospective study of the health of children in different ethnic groups, with particular reference to the effect of inbreeding. *European journal of human genetics : EJHG*. 1993;1(3):206-19. Epub 1993/01/01. PubMed PMID: 8044647.
140. Overall AD, Ahmad M, Thomas MG, Nichols RA. An analysis of consanguinity and social structure within the UK Asian population using microsatellite data. *Annals of human genetics*. 2003;67(Pt 6):525-37. Epub 2003/12/04. PubMed PMID: 14641240.
141. Bittles AH. A community genetics perspective on consanguineous marriage. *Community genetics*. 2008;11(6):324-30. Epub 2008/08/12. doi: 10.1159/000133304. PubMed PMID: 18690000.
142. Al-Awadi SA, Naguib KK, Moussa MA, Farag TI, Teebi AS, el-Khalifa MY. The effect of consanguineous marriages on reproductive wastage. *Clinical genetics*. 1986;29(5):384-8. Epub 1986/05/01. PubMed PMID: 3742845.
143. Abdulrazzaq YM, Bener A, al-Gazali LI, al-Khayat AI, Micallef R, Gaber T. A study of possible deleterious effects of consanguinity. *Clinical genetics*. 1997;51(3):167-73. Epub 1997/03/01. PubMed PMID: 9137881.
144. Subramanyan R, Joy J, Venugopalan P, Sapru A, al Khusaiby SM. Incidence and spectrum of congenital heart disease in Oman. *Annals of tropical paediatrics*. 2000;20(4):337-41. Epub 2001/02/24. PubMed PMID: 11219172.
145. Saha N, Hamad RE, Mohamed S. Inbreeding effects on reproductive outcome in a Sudanese population. *Hum Hered*. 1990;40(4):208-12. Epub 1990/01/01. PubMed PMID: 2379925.
146. al Husain M, al Bunyan M. Consanguineous marriages in a Saudi population and the effect of inbreeding on prenatal and postnatal mortality. *Annals of tropical paediatrics*. 1997;17(2):155-60. Epub 1997/06/01. PubMed PMID: 9230979.
147. Overall AD. The influence of the wahlund effect on the consanguinity hypothesis: consequences for recessive disease incidence in a socially structured pakistani population. *Hum Hered*. 2009;67(2):140-4. Epub 2008/12/17. doi: 10.1159/000179561. PubMed PMID: 19077430.
148. Nabulsi MM, Tamim H, Sabbagh M, Obeid MY, Yunis KA, Bitar FF. Parental consanguinity and congenital heart malformations in a developing country. *American journal of medical genetics Part A*. 2003;116A(4):342-7. Epub 2003/01/11. doi: 10.1002/ajmg.a.10020. PubMed PMID: 12522788.
149. Yunis K, Mumtaz G, Bitar F, Chamseddine F, Kassab M, Rashkidi J, et al. Consanguineous marriage and congenital heart defects: a case-control study in the neonatal period. *American journal of medical genetics Part A*. 2006;140(14):1524-30. Epub 2006/06/10. doi: 10.1002/ajmg.a.31309. PubMed PMID: 16763961.
150. Shami SA, Qaisar R, Bittles AH. Consanguinity and adult morbidity in Pakistan. *Lancet*. 1991;338(8772):954. Epub 1991/10/12. PubMed PMID: 1681304.

151. Denic S, Bener A. Consanguinity decreases risk of breast cancer--cervical cancer unaffected. *British journal of cancer*. 2001;85(11):1675-9. Epub 2001/12/18. doi: 10.1054/bjoc.2001.2131. PubMed PMID: 11742487; PubMed Central PMCID: PMCPMC2363968.
152. Ismail J, Jafar TH, Jafary FH, White F, Faruqui AM, Chaturvedi N. Risk factors for non-fatal myocardial infarction in young South Asian adults. *Heart (British Cardiac Society)*. 2004;90(3):259-63. Epub 2004/02/18. PubMed PMID: 14966040; PubMed Central PMCID: PMCPMC1768096.
153. Jaber L, Shohat T, Rotter JL, Shohat M. Consanguinity and common adult diseases in Israeli Arab communities. *American journal of medical genetics*. 1997;70(4):346-8. Epub 1997/06/27. PubMed PMID: 9182771.
154. Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, et al. Inbreeding and risk of late onset complex disease. *J Med Genet*. 2003;40(12):925-32. Epub 2003/12/20. PubMed PMID: 14684692; PubMed Central PMCID: PMCPMC1735350.
155. Rudan I, Skaric-Juric T, Smolej-Narancic N, Janicijevic B, Rudan D, Klaric IM, et al. Inbreeding and susceptibility to osteoporosis in Croatian island isolates. *Collegium antropologicum*. 2004;28(2):585-601. Epub 2005/01/26. PubMed PMID: 15666589.
156. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, et al. Inbreeding and the genetic complexity of human hypertension. *Genetics*. 2003;163(3):1011-21. Epub 2003/03/29. PubMed PMID: 12663539; PubMed Central PMCID: PMCPMC1462484.
157. Dictionaries O. Consanguineous: Oxford University Press; 2010 [26 April 2012]. Available from: <http://oxforddictionaries.com/definition/consanguineous>.
158. EDITOR T. CONSANGUINEOUS MARRIAGE: Subject Often Regarded by Unscientific Methods of Thought and Effects Misunderstood – Consanguinity in Itself Probably Has no Genetic Importance – The Hereditary Traits Are the Things To Be Considered – Marriage of Kin May Be Either Good or Bad in Effect. *Journal of Heredity*. 1916;7(8):343-6.
159. Alvarez G, Ceballos FC, Quinteiro C. The role of inbreeding in the extinction of a European royal dynasty. *PloS one*. 2009;4(4):e5174. Epub 2009/04/16. doi: 10.1371/journal.pone.0005174. PubMed PMID: 19367331; PubMed Central PMCID: PMCPMC2664480.
160. Paul DB, Spencer HG. "It's Ok, We're Not Cousins by Blood": The Cousin Marriage Controversy in Historical Perspective. *PLoS biology*. 2008;6(12):e320. doi: 10.1371/journal.pbio.0060320.
161. Davenport CB. *Heredity in Relation to Eugenics*: Henry Holt and Company; 1911.
162. Penrose CA. Sanitary conditions in the Bahama Islands. In: Shattuck GB, editor. [Baltimore: The Friedenwald company]; 1905.
163. Ottenheimer M. *Forbidden relatives: the American myth of cousin marriage*: University of Illinois Press; 1996.
164. Ingrao CW. *The Habsburg Monarchy, 1618-1815*: Cambridge University Press; 2000.

165. Middleton R. Brother-Sister and Father-Daughter Marriage in Ancient Egypt. *American Sociological Review*. 1962;27(5):603-11. doi: 10.2307/2089618.
166. Ager SL. Familiarity Breeds: Incest and the Ptolemaic Dynasty. *The Journal of Hellenic Studies*. 2005;125:1-34. doi: 10.2307/30033343.
167. Bittles AH, Mason WM, Greene J, Rao NA. Reproductive behavior and health in consanguineous marriages. *Science*. 1991;252(5007):789-94. Epub 1991/05/10. PubMed PMID: 2028254.
168. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, et al. Recent human effective population size estimated from linkage disequilibrium. *Genome research*. 2007;17(4):520-6. Epub 2007/03/14. doi: 10.1101/gr.6023607. PubMed PMID: 17351134; PubMed Central PMCID: PMC1832099.
169. Schull WJ. THE EFFECT OF CHRISTIANITY ON CONSANGUINITY IN NAGASAKI*. *American Anthropologist*. 1953;55(1):74-88. doi: 10.1525/aa.1953.55.1.02a00060.
170. Bittles AH. *Consanguinity in Context*: Cambridge University Press; 2012.
171. Cavalli-Sforza LL, Moroni A, Zei G. *Consanguinity, Inbreeding, and Genetic Drift in Italy (MPB-39)*: Princeton University Press; 2013.
172. de Costa C. Pregnancy outcomes in Lebanese-born women in western Sydney. *The Medical journal of Australia*. 1988;149(9):457-60. Epub 1988/11/07. PubMed PMID: 3185340.
173. Hussain R, Bittles AH. The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *Journal of biosocial science*. 1998;30(2):261-75. Epub 1998/09/25. PubMed PMID: 9746828.
174. Bittles AH. The role and significance of consanguinity as a demographic variable. *Population and development review*. 1994:561-84.
175. Modell B, Darr A. Science and society: genetic counselling and customary consanguineous marriage. *Nature reviews Genetics*. 2002;3(3):225-9. Epub 2002/04/25. doi: 10.1038/nrg754. PubMed PMID: 11972160.
176. Wahlund S. ZUSAMMENSETZUNG VON POPULATIONEN UND KORRELATIONSERSCHINUNGEN VOM STANDPUNKT DER VERERBUNGSLEHRE AUS BETRACHTET. *Hereditas*. 1928;11(1):65-106. doi: 10.1111/j.1601-5223.1928.tb02483.x.
177. Rittler M, Liascovich R, Lopez-Camelo J, Castilla EE. Parental consanguinity in specific types of congenital anomalies. *American journal of medical genetics*. 2001;102(1):36-43. Epub 2001/07/27. PubMed PMID: 11471170.
178. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. *American journal of human genetics*. 2008;83(3):359-72. Epub 2008/09/02. doi: 10.1016/j.ajhg.2008.08.007. PubMed PMID: 18760389; PubMed Central PMCID: PMC1832099.
179. Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS genetics*. 2009;5(3):e1000415. Epub 2009/03/14. doi: 10.1371/journal.pgen.1000415. PubMed PMID: 19282984; PubMed Central PMCID: PMC1832099.

180. Imaizumi Y. A recent survey of consanguineous marriages in Japan. *Clinical genetics*. 1986;30(3):230-3. Epub 1986/09/01. PubMed PMID: 3780039.
181. Bener A, Alali KA. Consanguineous marriage in a newly developed country: the Qatari population. *Journal of biosocial science*. 2006;38(2):239-46. Epub 2006/02/24. doi: 10.1017/s0021932004007060. PubMed PMID: 16490156.
182. al-Gazali LI, Bener A, Abdulrazzaq YM, Micallef R, al-Khayat AI, Gaber T. Consanguineous marriages in the United Arab Emirates. *Journal of biosocial science*. 1997;29(4):491-7. Epub 1999/01/09. PubMed PMID: 9881148.
183. Jurdi R, Saxena PC. The prevalence and correlates of consanguineous marriages in Yemen: similarities and contrasts with other Arab countries. *Journal of biosocial science*. 2003;35(1):1-13. Epub 2003/01/23. PubMed PMID: 12537152.
184. Wikipedia.org. Cousin marriage law in the United States by state 2013 [cited 2013 15-10-2013]. Available from: http://en.wikipedia.org/wiki/Cousin_marriage_law_in_the_United_States_by_state.
185. Hashmi MA. Frequency of consanguinity and its effect on congenital malformation--a hospital based study. *JPMA The Journal of the Pakistan Medical Association*. 1997;47(3):75-8. Epub 1997/03/01. PubMed PMID: 9131857.
186. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. *Nature genetics*. 2012;44(6):623-30. Epub 2012/05/30. doi: 10.1038/ng.2303. PubMed PMID: 22641211.
187. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*. 1980;32(3):314-31. Epub 1980/05/01. PubMed PMID: 6247908; PubMed Central PMCID: PMC1686077.
188. Brice A, Chalmers I. Medical journal editors and publication bias. *BMJ*. 2013;347:f6170. Epub 2013/10/24. doi: 10.1136/bmj.f6170. PubMed PMID: 24150668.
189. Blumenthal MN, Amos DB, Noreen H. Genetic mapping of Ir locus in man: linkage to second locus of HL-A. *Science*. 1974;184(4143):1301-3. Epub 1974/06/21. PubMed PMID: 4833283.
190. McIntosh I, Clough MV, Schaffer AA, Puffenberger EG, Horton VK, Peters K, et al. Fine mapping of the nail-patella syndrome locus at 9q34. *American journal of human genetics*. 1997;60(1):133-42. Epub 1997/01/01. PubMed PMID: 8981956; PubMed Central PMCID: PMC1712569.
191. Rahim NG, Harismendy O, Topol EJ, Frazer KA. Genetic determinants of phenotypic diversity in humans. *Genome biology*. 2008;9(4):215. Epub 2008/04/29. doi: 10.1186/gb-2008-9-4-215. PubMed PMID: 18439327; PubMed Central PMCID: PMC2643926.
192. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516-7. Epub 1996/09/13. PubMed PMID: 8801636.
193. International Multiple Sclerosis Genetics C, Wellcome Trust Case Control C, Sawcer S, Hellenthal G, Pirinen M, Spencer CC, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*. 2011;476(7359):214-9. Epub 2011/08/13. doi: 10.1038/nature10251. PubMed PMID: 21833088; PubMed Central PMCID: PMC3182531.

194. Soranzo N, Sanna S, Wheeler E, Gieger C, Radke D, Dupuis J, et al. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes*. 2010;59(12):3229-39. Epub 2010/09/23. doi: 10.2337/db10-0502. PubMed PMID: 20858683; PubMed Central PMCID: PMC2992787.
195. Diabetes Genetics Initiative of Broad Institute of H, Mit LU, Novartis Institutes of BioMedical R, Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007;316(5829):1331-6. Epub 2007/04/28. doi: 10.1126/science.1142358. PubMed PMID: 17463246.
196. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007;357(5):443-53. Epub 2007/07/20. doi: 10.1056/NEJMoa072366. PubMed PMID: 17634449; PubMed Central PMCID: PMC2719290.
197. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007;316(5826):889-94. Epub 2007/04/17. doi: 10.1126/science.1141634. PubMed PMID: 17434869; PubMed Central PMCID: PMC2646098.
198. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*. 2010;42(11):937-48. Epub 2010/10/12. doi: 10.1038/ng.686. PubMed PMID: 20935630; PubMed Central PMCID: PMC3014648.
199. Duncan EL, Danoy P, Kemp JP, Leo PJ, McCloskey E, Nicholson GC, et al. Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS genetics*. 2011;7(4):e1001372. Epub 2011/05/03. doi: 10.1371/journal.pgen.1001372. PubMed PMID: 21533022; PubMed Central PMCID: PMC3080863.
200. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, et al. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American journal of human genetics*. 2008;82(1):139-49. Epub 2008/01/09. doi: 10.1016/j.ajhg.2007.11.001. PubMed PMID: 18179892; PubMed Central PMCID: PMC2253977.
201. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010;466(7307):707-13. Epub 2010/08/06. doi: 10.1038/nature09270. PubMed PMID: 20686565; PubMed Central PMCID: PMC3039276.
202. Ellison JW, Rosenfeld JA, Shaffer LG. Genetic basis of intellectual disability. *Annual review of medicine*. 2013;64:441-50. Epub 2012/10/02. doi: 10.1146/annurev-med-042711-140053. PubMed PMID: 23020879.
203. Bergsagel DE. The chronic leukemias: a review of disease manifestations and the aims of therapy. *Canadian Medical Association journal*. 1967;96(25):1615-20. Epub 1967/06/24. PubMed PMID: 5338329; PubMed Central PMCID: PMC1923088.
204. Tkachuk DC, Westbrook CA, Andreeff M, Donlon TA, Cleary ML, Suryanarayan K, et al. Detection of bcr-abl fusion in chronic myelogenous leukemia

- by in situ hybridization. *Science*. 1990;250(4980):559-62. Epub 1990/10/26. PubMed PMID: 2237408.
205. Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*. 1989;86(8):2766-70. Epub 1989/04/01. PubMed PMID: 2565038; PubMed Central PMCID: PMC286999.
206. Liu WO, Oefner PJ, Qian C, Odom RS, Francke U. Denaturing HPLC-identified novel FBN1 mutations, polymorphisms, and sequence variants in Marfan syndrome and related connective tissue disorders. *Genetic testing*. 1997;1(4):237-42. Epub 1997/01/01. PubMed PMID: 10464652.
207. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*. 1986;51 Pt 1:263-73. Epub 1986/01/01. PubMed PMID: 3472723.
208. Consortium TGP. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-73. doi: <http://www.nature.com/nature/journal/v467/n7319/abs/10.1038-nature09534-unlocked.html#supplementary-information>.
209. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 2007;4(11):903-5. Epub 2007/10/16. doi: 10.1038/nmeth1111. PubMed PMID: 17934467.
210. Hodges E, Xuan Z, Baliya V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. *Nature genetics*. 2007;39(12):1522-7. Epub 2007/11/06. doi: 10.1038/ng.2007.42. PubMed PMID: 17982454.
211. Tewhey R, Nakano M, Wang X, Pabon-Pena C, Novak B, Giuffre A, et al. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome biology*. 2009;10(10):R116. Epub 2009/10/20. doi: 10.1186/gb-2009-10-10-r116. PubMed PMID: 19835619; PubMed Central PMCID: PMC2784331.
212. Ng SB. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009;461:272-6.
213. Shendure J. Next-generation human genetics. *Genome biology*. 2011;12(9):408. Epub 2011/09/17. doi: 10.1186/gb-2011-12-9-408. PubMed PMID: 21920048; PubMed Central PMCID: PMC3308046.
214. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-9. Epub 2012/05/19. doi: 10.1126/science.1219240. PubMed PMID: 22604720; PubMed Central PMCID: PMC3708544.
215. Vache C, Besnard T, le Berre P, Garcia-Garcia G, Baux D, Larrieu L, et al. Usher syndrome type 2 caused by activation of an USH2A pseudoexon: implications for diagnosis and therapy. *Human mutation*. 2012;33(1):104-8. Epub 2011/10/20. doi: 10.1002/humu.21634. PubMed PMID: 22009552.

216. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(45):19096-101. Epub 2009/10/29. doi: 10.1073/pnas.0910672106. PubMed PMID: 19861545; PubMed Central PMCID: PMC2768590.
217. Alsaadi MM, Gaunt TR, Boustred CR, Guthrie PA, Liu X, Lenzi L, et al. From a single whole exome read to notions of clinical screening: primary ciliary dyskinesia and RSPH9 p.Lys268del in the Arabian Peninsula. *Annals of human genetics*. 2012;76(3):211-20. Epub 2012/03/06. doi: 10.1111/j.1469-1809.2012.00704.x. PubMed PMID: 22384920.
218. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. Epub 2012/09/08. doi: 10.1038/nature11247. PubMed PMID: 22955616; PubMed Central PMCID: PMC3439153.
219. Alshammari MJ, Al-Otaibi L, Alkuraya FS. Mutation in RAB33B, which encodes a regulator of retrograde Golgi transport, defines a second Dyggve--Melchior--Clausen locus. *J Med Genet*. 2012;49(7):455-61. Epub 2012/06/02. doi: 10.1136/jmedgenet-2011-100666. PubMed PMID: 22652534.
220. Sailer A, Scholz SW, Gibbs JR, Tucci A, Johnson JO, Wood NW, et al. Exome sequencing in an SCA14 family demonstrates its utility in diagnosing heterogeneous diseases. *Neurology*. 2012;79(2):127-31. Epub 2012/06/08. doi: 10.1212/WNL.0b013e31825f048e. PubMed PMID: 22675081; PubMed Central PMCID: PMC3390538.
221. McGrath JA, Stone KL, Begum R, Simpson MA, Dopping-Hepenstal PJ, Liu L, et al. Germline Mutation in EXPH5 Implicates the Rab27B Effector Protein Slac2-b in Inherited Skin Fragility. *American journal of human genetics*. 2012;91(6):1115-21. Epub 2012/11/28. doi: 10.1016/j.ajhg.2012.10.012. PubMed PMID: 23176819; PubMed Central PMCID: PMC3516608.
222. Heller MJ. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*. 2002;4:129-53. Epub 2002/07/16. doi: 10.1146/annurev.bioeng.4.020702.153438. PubMed PMID: 12117754.
223. Alsolami R, Knight SJ, Schuh A. Clinical application of targeted and genome-wide technologies: can we predict treatment responses in chronic lymphocytic leukemia? *Personalized medicine*. 2013;10(4):361-76. Epub 2014/03/13. doi: 10.2217/pme.13.33. PubMed PMID: 24611071; PubMed Central PMCID: PMC3943176.
224. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) 2013 [cited 2013 15-10-2013]. Available from: www.genome.gov/sequencingcosts.
225. Min Jou W, Haegeman G, Ysebaert M, Fiers W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*. 1972;237(5350):82-8. Epub 1972/05/12. PubMed PMID: 4555447.
226. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*. 1977;74(2):560-4. Epub 1977/02/01. PubMed PMID: 265521; PubMed Central PMCID: PMC392330.

227. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977;74(12):5463-7. Epub 1977/12/01. PubMed PMID: 271968; PubMed Central PMCID: PMC431765.
228. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. *PLoS biology*. 2007;5(10):e254. Epub 2007/09/07. doi: 10.1371/journal.pbio.0050254. PubMed PMID: 17803354; PubMed Central PMCID: PMC1964779.
229. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. *BioEssays : news and reviews in molecular, cellular and developmental biology*. 2010;32(6):524-36. Epub 2010/05/21. doi: 10.1002/bies.200900181. PubMed PMID: 20486139.
230. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-80. Epub 2005/08/02. doi: 10.1038/nature03959. PubMed PMID: 16056220; PubMed Central PMCID: PMC1464427.
231. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005;309(5741):1728-32. Epub 2005/08/06. doi: 10.1126/science.1117389. PubMed PMID: 16081699.
232. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-9. Epub 2008/11/07. doi: 10.1038/nature07517. PubMed PMID: 18987734; PubMed Central PMCID: PMC2581791.
233. Glenn TC. Field guide to next-generation DNA sequencers. *Molecular ecology resources*. 2011;11(5):759-69. Epub 2011/05/20. doi: 10.1111/j.1755-0998.2011.03024.x. PubMed PMID: 21592312.
234. Geohive.com. Current world population (ranked) 2013 [cited 2013 15-10-13]. Available from: http://www.geohive.com/earth/population_now.aspx.
235. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. Epub 2012/11/07. doi: 10.1038/nature11632. PubMed PMID: 23128226; PubMed Central PMCID: PMCPmc3498066.
236. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research*. 1999;9(8):677-9. Epub 1999/08/14. PubMed PMID: 10447503.
237. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001;29(1):308-11. Epub 2000/01/11. PubMed PMID: 11125122; PubMed Central PMCID: PMC29783.
238. Kitts A SS. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation National Center for Biotechnology Information (US)2002 [updated Updated 2011 Feb 2]. Chapter 5:[Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21088/>].
239. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in

- genotyping and haplotyping studies. *Human mutation*. 2010;31(1):67-73. Epub 2009/10/31. doi: 10.1002/humu.21137. PubMed PMID: 19877174; PubMed Central PMCID: PMC2797835.
240. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. *Cell*. 2011;147(1):32-43. Epub 2011/10/04. doi: 10.1016/j.cell.2011.09.008. PubMed PMID: 21962505; PubMed Central PMCID: PMC3656718.
241. Abbott A. Rare-disease project has global ambitions. *Nature*. 2011;472(7341):17. Epub 2011/04/09. doi: 10.1038/472017a. PubMed PMID: 21475168.
242. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*. 2013;34(1):57-65. Epub 2012/10/04. doi: 10.1002/humu.22225. PubMed PMID: 23033316; PubMed Central PMCID: PMCPmc3558800.
243. Ng P, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*. 2003;31:3812 - 4. PubMed PMID: doi:10.1093/nar/gkg509.
244. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Meth*. 2010;7(4):248-9. doi: http://www.nature.com/nmeth/journal/v7/n4/supinfo/nmeth0410-248_S1.html.
245. Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, et al. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic acids research*. 1989;17(7):2503-16. Epub 1989/04/11. PubMed PMID: 2785681; PubMed Central PMCID: PMCPmc317639.
246. Day IN, Humphries SE. Electrophoresis for genotyping: microtiter array diagonal gel electrophoresis on horizontal polyacrylamide gels, hydrolink, or agarose. *Analytical biochemistry*. 1994;222(2):389-95. Epub 1994/11/01. doi: 10.1006/abio.1994.1507. PubMed PMID: 7864363.
247. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-60. doi: 10.1093/bioinformatics/btp324.
248. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. *Genome research*. 2009;19:1124 - 32. PubMed PMID: doi:10.1101/gr.088013.108.
249. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 2010;20:1297 - 303. PubMed PMID: doi:10.1101/gr.107524.110.
250. Li H. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078-9.
251. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8. Epub

- 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMC3137218.
252. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069-70. doi: 10.1093/bioinformatics/btq330.
253. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010;38(16):e164. Epub 2010/07/06. doi: 10.1093/nar/gkq603. PubMed PMID: 20601685; PubMed Central PMCID: PMC2938201.
254. Boeva V, Popova T, Bleakley K, Chiche P, Cappelletti J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423-5. Epub 2011/12/14. doi: 10.1093/bioinformatics/btr670. PubMed PMID: 22155870; PubMed Central PMCID: PMC3268243.
255. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*. 2011;32:358-68.
256. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997;25(17):3389-402. Epub 1997/09/01. PubMed PMID: 9254694; PubMed Central PMCID: PMC146917.
257. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American journal of human genetics*. 2011;88(4):440-9. Epub 2011/04/05. doi: 10.1016/j.ajhg.2011.03.004. PubMed PMID: 21457909; PubMed Central PMCID: PMC3071923.
258. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*. 2013;41(Database issue):D808-15. Epub 2012/12/04. doi: 10.1093/nar/gks1094. PubMed PMID: 23203871; PubMed Central PMCID: PMC3531103.
259. Inglis PN, Boroevich KA, Leroux MR. Piecing together a ciliome. *Trends in genetics : TIG*. 2006;22(9):491-500. Epub 2006/07/25. doi: 10.1016/j.tig.2006.07.006. PubMed PMID: 16860433.
260. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*. 2004;32(Web Server issue):W526-31. Epub 2004/06/25. doi: 10.1093/nar/gkh468. PubMed PMID: 15215442; PubMed Central PMCID: PMC1441606.
261. Wright S. Coefficients of Inbreeding and Relationship. *The American Naturalist*. 1922;(56):330-8.
262. Metzker ML. Sequencing technologies - the next generation. *Nature reviews Genetics*. 2010;11(1):31-46. Epub 2009/12/10. doi: 10.1038/nrg2626. PubMed PMID: 19997069.
263. Bonetta L. Whole-Genome Sequencing Breaks the Cost Barrier. *Cell*. 2010;141(6):917-9. doi: 10.1016/j.cell.2010.05.034.
264. Pettersson E, Lundberg J, Ahmadian A. Generations of sequencing technologies. *Genomics*. 2009;93(2):105-11. doi: 10.1016/j.ygeno.2008.10.003.

265. Hedges DJ. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PloS one*. 2011;6:e18595.
266. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*. 2010;19(R2):R145-R51. doi: 10.1093/hmg/ddq333.
267. Bick D, Dimmock D. Whole exome and whole genome sequencing. *Current opinion in pediatrics*. 2011;23(6):594-600. Epub 2011/09/02. doi: 10.1097/MOP.0b013e32834b20ec. PubMed PMID: 21881504.
268. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306(5696):636-40. Epub 2004/10/23. doi: 10.1126/science.1105136. PubMed PMID: 15499007.
269. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM. A general framework for estimating the relative pathogenicity of human genetic variants. 2014;46(3):310-5. doi: 10.1038/ng.2892. PubMed PMID: 24487276.
270. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015. Epub 2015/01/15. doi: 10.1093/bioinformatics/btv009. PubMed PMID: 25583119.
271. Ritchie GR, Dunham I. Functional annotation of noncoding sequence variants. 2014;11(3):294-6. doi: 10.1038/nmeth.2832. PubMed PMID: 24487584.
272. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. *Human genetics*. 2011;129(4):351-70. Epub 2011/02/19. doi: 10.1007/s00439-011-0964-2. PubMed PMID: 21331778.
273. Brandstätter A, Sängner T, Lutz-Bonengel S, Parson W, Béraud-Colomb E, Wen B, et al. Phantom mutation hotspots in human mitochondrial DNA. *ELECTROPHORESIS*. 2005;26(18):3414-29. doi: 10.1002/elps.200500307.
274. Castleman VH, Romio L, Chodhari R, Hirst RA, de Castro SC, Parker KA, et al. Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *American journal of human genetics*. 2009;84(2):197-209. Epub 2009/02/10. doi: 10.1016/j.ajhg.2009.01.011. PubMed PMID: 19200523; PubMed Central PMCID: PMC2668031.
275. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech*. 2009;27(2):182-9. doi: http://www.nature.com/nbt/journal/v27/n2/supinfo/nbt.1523_S1.html.
276. Bainbridge M, Wang M, Burgess D, Kovar C, Rodesch M, D'Ascenzo M, et al. Whole exome capture in solution with 3 Gbp of data. *Genome biology*. 2010;11(6):R62. PubMed PMID: doi:10.1186/gb-2010-11-6-r62.
277. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome biology*. 2011;12(9):R94. Epub 2011/10/01. doi: 10.1186/gb-2011-12-9-r94. PubMed PMID: 21955854; PubMed Central PMCID: PMC3308057.
278. Chan EY. Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery

- Single Nucleotide Polymorphisms. In: Komar AA, editor. *Methods in Molecular Biology*. 578: Humana Press; 2009. p. 95-111.
279. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-71. Epub 2009/06/30. doi: 10.1093/bioinformatics/btp394. PubMed PMID: 19561018; PubMed Central PMCID: PMC2781750.
280. Homer N, Merriman B, Nelson S. Local alignment of two-base encoded DNA sequence. *BMC Bioinformatics*. 2009;10(1):175. PubMed PMID: doi:10.1186/1471-2105-10-175.
281. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012;9(4):357-9. doi: <http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html#supplementary-information>.
282. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*. 2008;18:1851 - 8. PubMed PMID: doi:10.1101/gr.078212.108.
283. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966-7. doi: 10.1093/bioinformatics/btp336.
284. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*. 2013;14(2):178-92. Epub 2012/04/21. doi: 10.1093/bib/bbs017. PubMed PMID: 22517427; PubMed Central PMCID: PMC3603213.
285. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. *PloS one*. 2013;8(9):e75619. Epub 2013/10/03. doi: 10.1371/journal.pone.0075619. PubMed PMID: 24086590; PubMed Central PMCID: PMC3785481.
286. Le S, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research*. 2011;21:952 - 60. PubMed PMID: doi:10.1101/gr.113084.110.
287. Quinlan A, Stewart D, Stromberg M, Marth G. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods*. 2008;5:179 - 81. PubMed PMID: doi:10.1038/nmeth.1172.
288. Li D, Guo Y, Shao H, Tellier L, Wang J, Xiang Z, et al. Genetic diversity, molecular phylogeny and selection evidence of the silkworm mitochondria implicated by complete resequencing of 41 genomes. *BMC Evolutionary Biology*. 2010;10(1):81. PubMed PMID: doi:10.1186/1471-2148-10-81.
289. Li S, Wang S, Deng Q, Zheng A, Zhu J, Liu H, et al. Identification of Genome-Wide Variations among Three Elite Restorer Lines for Hybrid-Rice. *PloS one*. 2012;7(2):e30952. doi: 10.1371/journal.pone.0030952.
290. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012;13:8. Epub 2012/01/14. doi: 10.1186/1471-2105-13-8. PubMed PMID: 22239737; PubMed Central PMCID: PMC3292476.

291. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*. 2007;81(5):1084-97. Epub 2007/10/10. doi: 10.1086/521987. PubMed PMID: 17924348; PubMed Central PMCID: PMC2265661.
292. Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomes, Genetics*. 2011;1(6):457-70. doi: 10.1534/g3.111.001198.
293. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010;34(8):816-34. Epub 2010/11/09. doi: 10.1002/gepi.20533. PubMed PMID: 21058334; PubMed Central PMCID: PMC3175618.
294. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007;81:559 - 75. PubMed PMID: doi:10.1086/519795.
295. Woods CG, Valente EM, Bond J, Roberts E. A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J Med Genet*. 2004;41(8):e101. Epub 2004/08/03. doi: 10.1136/jmg.2003.016873. PubMed PMID: 15286161; PubMed Central PMCID: PMC1735872.
296. Carr IM, Bhaskar S, O'Sullivan J, Aldahmesh MA, Shamseldin HE, Markham AF, et al. Autozygosity mapping with exome sequence data. *Human mutation*. 2013;34(1):50-6. Epub 2012/10/24. doi: 10.1002/humu.22220. PubMed PMID: 23090942.
297. Carr IM, Flintoff KJ, Taylor GR, Markham AF, Bonthron DT. Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. *Human mutation*. 2006;27(10):1041-6. Epub 2006/08/31. doi: 10.1002/humu.20383. PubMed PMID: 16941472.
298. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics / editorial board, Andreas D Baxeavanis [et al]*. 2012;Chapter 1:Unit1 13. Epub 2012/09/06. doi: 10.1002/0471250953.bi0113s39. PubMed PMID: 22948725.
299. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*. 2014;42(Database issue):D980-5. Epub 2013/11/16. doi: 10.1093/nar/gkt1113. PubMed PMID: 24234437; PubMed Central PMCID: PMCPMC3965032.
300. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. *Human mutation*. 2011;32(5):557-63. Epub 2011/04/27. doi: 10.1002/humu.21438. PubMed PMID: 21520333.
301. McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome medicine*. 2014;6(3):26. Epub 2014/06/20. doi: 10.1186/gm543. PubMed PMID: 24944579; PubMed Central PMCID: PMCPmc4062061.

302. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics*. 2006;7:61-80. Epub 2006/07/11. doi: 10.1146/annurev.genom.7.080505.115630. PubMed PMID: 16824020.
303. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*. 2009;25(21):2744-50. doi: 10.1093/bioinformatics/btp528.
304. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*. 2009;37(9):e67. Epub 2009/04/03. doi: 10.1093/nar/gkp215. PubMed PMID: 19339519; PubMed Central PMCID: PMC2685110.
305. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069-70. Epub 2010/06/22. doi: 10.1093/bioinformatics/btq330. PubMed PMID: 20562413; PubMed Central PMCID: PMC2916720.
306. Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*. 2013;29(12):1504-10. Epub 2013/04/27. doi: 10.1093/bioinformatics/btt182. PubMed PMID: 23620363; PubMed Central PMCID: PMC3673218.
307. Capriotti E, Altman RB. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics*. 2011;98(4):310-7. Epub 2011/07/19. doi: 10.1016/j.ygeno.2011.06.010. PubMed PMID: 21763417; PubMed Central PMCID: PMC3371640.
308. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*. 2012;40(W1):W452-W7. doi: 10.1093/nar/gks539.
309. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protocols*. 2009;4(8):1073-81.
310. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol*. 2010;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025.
311. Cooper G, Goode D, Ng S, Sidow A, Bamshad M, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods*. 2010;7:250 - 1. PubMed PMID: doi:10.1038/nmeth0410-250.
312. Cooper GM. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15:901-13.
313. Pollard K, Hubisz M, Rosenbloom K, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*. 2010;20:110 - 21. PubMed PMID: doi:10.1101/gr.097857.109.
314. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research*. 2007;35:3823 - 35. PubMed PMID: doi:10.1093/nar/gkm238.

315. Conde L, Vaquerizas J, Dopazo H, Arbiza L, Reumers J, Rousseau F, et al. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic acids research*. 2006;34:W621 - 5. PubMed PMID: doi:10.1093/nar/gkl071.
316. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F. SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic acids research*. 2005;33(suppl 1):D527-D32. doi: 10.1093/nar/gki086.
317. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*. 2011;39(17):e118. doi: 10.1093/nar/gkr407.
318. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*. 2003;31(1):334-41. doi: 10.1093/nar/gkg115.
319. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research*. 2010;38(suppl 1):D204-D10. doi: 10.1093/nar/gkp1019.
320. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human mutation*. 2009;30(8):1237-44. doi: 10.1002/humu.21047.
321. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic acids research*. 2005;33(Web Server issue):W480-2. Epub 2005/06/28. doi: 10.1093/nar/gki372. PubMed PMID: 15980516; PubMed Central PMCID: PMC1160133.
322. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*. 2006;22(22):2729-34. doi: 10.1093/bioinformatics/btl423.
323. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic acids research*. 2002;30(17):3894-900. doi: 10.1093/nar/gkf493.
324. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*. 2005;21(14):3176-8. doi: 10.1093/bioinformatics/bti486.
325. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nature reviews Genetics*. 2011;12(10):683-91. doi: http://www.nature.com/nrg/journal/v12/n10/supinfo/nrg3051_S1.html.
326. Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. Identification of deleterious synonymous variants in human genomes. *Bioinformatics*. 2013;29(15):1843-50. Epub 2013/06/06. doi: 10.1093/bioinformatics/btt308. PubMed PMID: 23736532.
327. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*. 2000;28(1):27-30. Epub 1999/12/11. PubMed PMID: 10592173; PubMed Central PMCID: PMCPmc102409.

328. Yue P, Melamud E, Moulton J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*. 2006;7:166. Epub 2006/03/23. doi: 10.1186/1471-2105-7-166. PubMed PMID: 16551372; PubMed Central PMCID: PMC1435944.
329. Marques-Pinheiro A, Marduel M, Rabes J-P, Devillers M, Villeger L, Allard D, et al. A fourth locus for autosomal dominant hypercholesterolemia maps at 16q22.1. *European journal of human genetics : EJHG*. 2010;18(11):1236-42. doi: <http://www.nature.com/ejhg/journal/v18/n11/supinfo/ejhg201094s1.html>.
330. Audrézet M-P, Chen J-M, Raguénès O, Chuzhanova N, Giteau K, Maréchal CL, et al. Genomic rearrangements in the CFTR gene: Extensive allelic heterogeneity and diverse mutational mechanisms. *Human mutation*. 2004;23(4):343-57. doi: 10.1002/humu.20009.
331. Zheng XL, Sadler JE. Pathogenesis of Thrombotic Microangiopathies. *Annual Review of Pathology: Mechanisms of Disease*. 2008;3(1):249-77. doi: doi:10.1146/annurev.pathmechdis.3.121806.154311.
332. Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS genetics*. 2010;6(6):e1000991. Epub 2010/06/26. doi: 10.1371/journal.pgen.1000991. PubMed PMID: 20577567; PubMed Central PMCID: PMC2887469.
333. Norio R. Finnish Disease Heritage I: characteristics, causes, background. *Human genetics*. 2003;112(5-6):441-56. Epub 2003/03/11. doi: 10.1007/s00439-002-0875-3. PubMed PMID: 12627295.
334. Norio R. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Human genetics*. 2003;112(5-6):457-69. Epub 2003/03/11. doi: 10.1007/s00439-002-0876-2. PubMed PMID: 12627296.
335. Norio R. The Finnish Disease Heritage III: the individual diseases. *Human genetics*. 2003;112(5-6):470-526. Epub 2003/03/11. doi: 10.1007/s00439-002-0877-1. PubMed PMID: 12627297.
336. Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, et al. Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *American journal of human genetics*. 2006;78(5):889-96. Epub 2006/04/28. doi: 10.1086/503875. PubMed PMID: 16642444; PubMed Central PMCID: PMC1474039.
337. Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D. Phasing of many thousands of genotyped samples. *American journal of human genetics*. 2012;91(2):238-51. Epub 2012/08/14. doi: 10.1016/j.ajhg.2012.06.013. PubMed PMID: 22883141; PubMed Central PMCID: PMC3415548.
338. Blumenthal MN. Genetic, epigenetic, and environmental factors in asthma and allergy. *Annals of Allergy, Asthma & Immunology*. 2012;108(2):69-73. doi: 10.1016/j.anai.2011.12.003.
339. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikäinen LP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*. 2012;44(3):269-76. Epub 2012/01/31. doi: 10.1038/ng.1073. PubMed PMID: 22286219; PubMed Central PMCID: PMC3605033.

340. Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ. Epistasis in sporadic Alzheimer's disease. *Neurobiology of Aging*. 2009;30(9):1333-49. doi: 10.1016/j.neurobiolaging.2007.11.027.
341. Farooqi S, Rau H, Whitehead J, O'Rahilly S. ob gene mutations and human obesity. *The Proceedings of the Nutrition Society*. 1998;57(3):471-5. Epub 1998/10/30. PubMed PMID: 9794006.
342. Lucas JS, Adam EC, Goggin PM, Jackson CL, Powles-Glover N, Patel SH, et al. Static respiratory cilia associated with mutations in Dnahc11/DNAH11: a mouse model of PCD. *Human mutation*. 2012;33(3):495-503. Epub 2011/11/22. doi: 10.1002/humu.22001. PubMed PMID: 22102620.
343. Dutcher SK. Flagellar assembly in two hundred and fifty easy-to-follow steps. *Trends in genetics : TIG*. 1995;11(10):398-404. Epub 1995/10/01. PubMed PMID: 7482766.
344. Gherman A, Davis EE, Katsanis N. The ciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia. *Nature genetics*. 2006;38(9):961-2. Epub 2006/08/31. doi: 10.1038/ng0906-961. PubMed PMID: 16940995.
345. Ostrowski LE, Blackburn K, Radde KM, Moyer MB, Schlatzer DM, Moseley A, et al. A proteomic analysis of human cilia: identification of novel components. *Molecular & cellular proteomics : MCP*. 2002;1(6):451-65. Epub 2002/08/10. PubMed PMID: 12169685.
346. Storm van's Gravesande K, Omran H. Primary ciliary dyskinesia: clinical presentation, diagnosis and genetics. *Annals of medicine*. 2005;37(6):439-49. Epub 2005/10/06. doi: 10.1080/07853890510011985. PubMed PMID: 16203616.
347. Nonaka S, Tanaka Y, Okada Y, Takeda S, Harada A, Kanai Y, et al. Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell*. 1998;95(6):829-37. Epub 1998/12/29. PubMed PMID: 9865700.
348. Ibanez-Tallon I, Pagenstecher A, Fliegauf M, Olbrich H, Kispert A, Ketelsen UP, et al. Dysfunction of axonemal dynein heavy chain Mdnah5 inhibits ependymal flow and reveals a novel mechanism for hydrocephalus formation. *Hum Mol Genet*. 2004;13(18):2133-41. Epub 2004/07/23. doi: 10.1093/hmg/ddh219. PubMed PMID: 15269178.
349. Zariwala MA, Knowles MR, Omran H. Genetic defects in ciliary structure and function. *Annual review of physiology*. 2007;69:423-50. Epub 2006/10/25. doi: 10.1146/annurev.physiol.69.040705.141301. PubMed PMID: 17059358.
350. Leigh MW, Pittman JE, Carson JL, Ferkol TW, Dell SD, Davis SD, et al. Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. *Genet Med*. 2009;11(7):473-87. Epub 2009/07/17. doi: 10.1097/GIM.0b013e3181a53562. PubMed PMID: 19606528; PubMed Central PMCID: PMC3739704.
351. Barbato A, Frischer T, Kuehni CE, Snijders D, Azevedo I, Baktai G, et al. Primary ciliary dyskinesia: a consensus statement on diagnostic and treatment approaches in children. *The European respiratory journal*. 2009;34(6):1264-76. Epub 2009/12/02. doi: 10.1183/09031936.00176608. PubMed PMID: 19948909.

352. Afzelius BA. The immotile-cilia syndrome: a microtubule-associated defect. *CRC critical reviews in biochemistry*. 1985;19(1):63-87. Epub 1985/01/01. PubMed PMID: 3907978.
353. Noone PG, Leigh MW, Sannuti A, Minnix SL, Carson JL, Hazucha M, et al. Primary ciliary dyskinesia: diagnostic and phenotypic features. *American journal of respiratory and critical care medicine*. 2004;169(4):459-67. Epub 2003/12/06. doi: 10.1164/rccm.200303-365OC. PubMed PMID: 14656747.
354. Torgersen J. Transposition of viscera, bronchiectasis and nasal polyps; a genetical analysis and a contribution to the problem of constitution. *Acta radiologica*. 1947;28(1):17-24. Epub 1947/02/28. PubMed PMID: 20295650.
355. Afzelius BA, Stenram U. Prevalence and genetics of immotile-cilia syndrome and left-handedness. *The International journal of developmental biology*. 2006;50(6):571-3. Epub 2006/06/03. doi: 10.1387/ijdb.052132ba. PubMed PMID: 16741872.
356. Katsuhara K, Kawamoto S, Wakabayashi T, Belsky JL. Situs inversus totalis and Kartagener's syndrome in a Japanese population. *Chest*. 1972;61(1):56-61. Epub 1972/01/01. PubMed PMID: 4538074.
357. Bush A, Chodhari R, Collins N, Copeland F, Hall P, Harcourt J, et al. Primary ciliary dyskinesia: current state of the art. *Archives of disease in childhood*. 2007;92(12):1136-40. Epub 2007/07/20. doi: 10.1136/adc.2006.096958. PubMed PMID: 17634184; PubMed Central PMCID: PMC2066071.
358. Rott HD. Kartagener's syndrome and the syndrome of immotile cilia. *Human genetics*. 1979;46(3):249-61. Epub 1979/02/15. PubMed PMID: 155641.
359. Reish O, Slatkin M, Chapman-Shimshoni D, Elizur A, Chioza B, Castleman V, et al. Founder mutation(s) in the RSPH9 gene leading to primary ciliary dyskinesia in two inbred Bedouin families. *Annals of human genetics*. 2010;74(2):117-25. Epub 2010/01/15. doi: 10.1111/j.1469-1809.2009.00559.x. PubMed PMID: 20070851; PubMed Central PMCID: PMC2853723.
360. Kennedy MP, Omran H, Leigh MW, Dell S, Morgan L, Molina PL, et al. Congenital heart disease and other heterotaxic defects in a large cohort of patients with primary ciliary dyskinesia. *Circulation*. 2007;115(22):2814-21. Epub 2007/05/23. doi: 10.1161/circulationaha.106.649038. PubMed PMID: 17515466.
361. Jeganathan D, Chodhari R, Meeks M, Faeroe O, Smyth D, Nielsen K, et al. Loci for primary ciliary dyskinesia map to chromosome 16p12.1-12.2 and 15q13.1-15.1 in Faroe Islands and Israeli Druze genetic isolates. *J Med Genet*. 2004;41(3):233-40. Epub 2004/02/27. PubMed PMID: 14985390; PubMed Central PMCID: PMC1735711.
362. Chilvers MA, Rutman A, O'Callaghan C. Ciliary beat pattern is associated with specific ultrastructural defects in primary ciliary dyskinesia. *The Journal of allergy and clinical immunology*. 2003;112(3):518-24. Epub 2003/09/19. PubMed PMID: 13679810.
363. O'Callaghan C. Innate pulmonary immunity: cilia. *Pediatric pulmonology Supplement*. 2004;26:72-3. Epub 2004/03/20. PubMed PMID: 15029603.
364. Satir P, Christensen ST. Overview of structure and function of mammalian cilia. *Annual review of physiology*. 2007;69:377-400. Epub 2006/10/03. doi: 10.1146/annurev.physiol.69.040705.141236. PubMed PMID: 17009929.

365. Yang P, Diener DR, Yang C, Kohno T, Pazour GJ, Dienes JM, et al. Radial spoke proteins of Chlamydomonas flagella. *Journal of cell science*. 2006;119(Pt 6):1165-74. Epub 2006/03/02. doi: 10.1242/jcs.02811. PubMed PMID: 16507594; PubMed Central PMCID: PMC1973137.
366. Porter ME, Sale WS. The 9 + 2 axoneme anchors multiple inner arm dyneins and a network of kinases and phosphatases that control motility. *The Journal of cell biology*. 2000;151(5):F37-42. Epub 2000/11/22. PubMed PMID: 11086017; PubMed Central PMCID: PMC2174360.
367. Smith EF, Yang P. The radial spokes and central apparatus: mechano-chemical transducers that regulate flagellar motility. *Cell motility and the cytoskeleton*. 2004;57(1):8-17. Epub 2003/12/03. doi: 10.1002/cm.10155. PubMed PMID: 14648553; PubMed Central PMCID: PMC1950942.
368. Busquets RM, Caballero-Rabasco MA, Velasco M, Lloreta J, Garcia-Algar O. Primary ciliary dyskinesia: clinical criteria indicating ultrastructural studies. *Archivos de bronconeumologia*. 2013;49(3):99-104. Epub 2012/12/26. doi: 10.1016/j.arbres.2012.10.007. PubMed PMID: 23265970.
369. Blouin JL, Meeks M, Radhakrishna U, Sainsbury A, Gehring C, Sail GD, et al. Primary ciliary dyskinesia: a genome-wide linkage analysis reveals extensive locus heterogeneity. *European journal of human genetics : EJHG*. 2000;8(2):109-18. Epub 2000/04/11. doi: 10.1038/sj.ejhg.5200429. PubMed PMID: 10757642.
370. Zariwala MA, Knowles MR, Leigh MW. Primary Ciliary Dyskinesia. In: Pagon RA, Adam MP, Bird TD, Dolan CR, Fong CT, Stephens K, editors. *GeneReviews*. Seattle WA: University of Washington, Seattle; 1993.
371. Olbrich H, Haffner K, Kispert A, Volkel A, Volz A, Sasmaz G, et al. Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nature genetics*. 2002;30(2):143-4. Epub 2002/01/15. doi: 10.1038/ng817. PubMed PMID: 11788826.
372. Hornef N, Olbrich H, Horvath J, Zariwala MA, Fliegauf M, Loges NT, et al. DNAH5 mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. *American journal of respiratory and critical care medicine*. 2006;174(2):120-6. Epub 2006/04/22. doi: 10.1164/rccm.200601-084OC. PubMed PMID: 16627867; PubMed Central PMCID: PMC2662904.
373. Faily M, Bartoloni L, Letourneau A, Munoz A, Falconnet E, Rossier C, et al. Mutations in DNAH5 account for only 15% of a non-preselected cohort of patients with primary ciliary dyskinesia. *J Med Genet*. 2009;46(4):281-6. Epub 2009/04/10. doi: 10.1136/jmg.2008.061176. PubMed PMID: 19357118.
374. Faily M, Saitta A, Munoz A, Falconnet E, Rossier C, Santamaria F, et al. DNAI1 mutations explain only 2% of primary ciliary dyskinesia. *Respiration; international review of thoracic diseases*. 2008;76(2):198-204. Epub 2008/04/25. doi: 10.1159/000128567. PubMed PMID: 18434704.
375. Zariwala MA, Leigh MW, Ceppa F, Kennedy MP, Noone PG, Carson JL, et al. Mutations of DNAI1 in primary ciliary dyskinesia: evidence of founder effect in a common mutation. *American journal of respiratory and critical care medicine*. 2006;174(8):858-66. Epub 2006/07/22. doi: 10.1164/rccm.200603-370OC. PubMed PMID: 16858015; PubMed Central PMCID: PMC2648054.

376. Pennarun G, Escudier E, Chapelin C, Bridoux AM, Cacheux V, Roger G, et al. Loss-of-function mutations in a human gene related to *Chlamydomonas reinhardtii* dynein IC78 result in primary ciliary dyskinesia. *American journal of human genetics*. 1999;65(6):1508-19. Epub 1999/12/01. doi: 10.1086/302683. PubMed PMID: 10577904; PubMed Central PMCID: PMC1288361.
377. Duquesnoy P, Escudier E, Vincensini L, Freshour J, Bridoux AM, Coste A, et al. Loss-of-function mutations in the human ortholog of *Chlamydomonas reinhardtii* ODA7 disrupt dynein arm assembly and cause primary ciliary dyskinesia. *American journal of human genetics*. 2009;85(6):890-6. Epub 2009/12/01. doi: 10.1016/j.ajhg.2009.11.008. PubMed PMID: 19944405; PubMed Central PMCID: PMC2790569.
378. Merveille AC, Davis EE, Becker-Heck A, Legendre M, Amirav I, Bataille G, et al. CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nature genetics*. 2011;43(1):72-8. Epub 2010/12/07. doi: 10.1038/ng.726. PubMed PMID: 21131972; PubMed Central PMCID: PMC3509786.
379. Blanchon S, Legendre M, Copin B, Duquesnoy P, Montantin G, Kott E, et al. Delineation of CCDC39/CCDC40 mutation spectrum and associated phenotypes in primary ciliary dyskinesia. *J Med Genet*. 2012;49(6):410-6. Epub 2012/06/14. doi: 10.1136/jmedgenet-2012-100867. PubMed PMID: 22693285.
380. Becker-Heck A, Zohn IE, Okabe N, Pollock A, Lenhart KB, Sullivan-Brown J, et al. The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation. *Nature genetics*. 2011;43(1):79-84. Epub 2010/12/07. doi: 10.1038/ng.727. PubMed PMID: 21131974; PubMed Central PMCID: PMC3132183.
381. Knowles MR, Leigh MW, Carson JL, Davis SD, Dell SD, Ferkol TW, et al. Mutations of DNAH11 in patients with primary ciliary dyskinesia with normal ciliary ultrastructure. *Thorax*. 2012;67(5):433-41. Epub 2011/12/21. doi: 10.1136/thoraxjnl-2011-200301. PubMed PMID: 22184204; PubMed Central PMCID: PMC3739700.
382. Loges NT, Olbrich H, Fenske L, Mussaffi H, Horvath J, Fliegauf M, et al. DNAI2 mutations cause primary ciliary dyskinesia with defects in the outer dynein arm. *American journal of human genetics*. 2008;83(5):547-58. Epub 2008/10/28. doi: 10.1016/j.ajhg.2008.10.001. PubMed PMID: 18950741; PubMed Central PMCID: PMC2668028.
383. Kott E, Duquesnoy P, Copin B, Legendre M, Dastot-Le Moal F, Montantin G, et al. Loss-of-function mutations in LRRC6, a gene essential for proper axonemal assembly of inner and outer dynein arms, cause primary ciliary dyskinesia. *American journal of human genetics*. 2012;91(5):958-64. Epub 2012/11/06. doi: 10.1016/j.ajhg.2012.10.003. PubMed PMID: 23122589; PubMed Central PMCID: PMC3487148.
384. Omran H, Kobayashi D, Olbrich H, Tsukahara T, Loges NT, Hagiwara H, et al. Ktu/PF13 is required for cytoplasmic pre-assembly of axonemal dyneins. *Nature*. 2008;456(7222):611-6. Epub 2008/12/05. doi: 10.1038/nature07471. PubMed PMID: 19052621; PubMed Central PMCID: PMC3279746.

385. Duriez B, Duquesnoy P, Escudier E, Bridoux AM, Escalier D, Rayet I, et al. A common variant in combination with a nonsense mutation in a member of the thioredoxin family causes primary ciliary dyskinesia. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(9):3336-41. Epub 2007/03/16. doi: 10.1073/pnas.0611405104. PubMed PMID: 17360648; PubMed Central PMCID: PMC1805560.
386. Panizzi JR, Becker-Heck A, Castleman VH, Al-Mutairi DA, Liu Y, Loges NT, et al. CCDC103 mutations cause primary ciliary dyskinesia by disrupting assembly of ciliary dynein arms. *Nature genetics*. 2012;44(6):714-9. Epub 2012/05/15. doi: 10.1038/ng.2277. PubMed PMID: 22581229; PubMed Central PMCID: PMC3371652.
387. Knowles MR, Leigh MW, Ostrowski LE, Huang L, Carson JL, Hazucha MJ, et al. Exome sequencing identifies mutations in CCDC114 as a cause of primary ciliary dyskinesia. *American journal of human genetics*. 2013;92(1):99-106. Epub 2012/12/25. doi: 10.1016/j.ajhg.2012.11.003. PubMed PMID: 23261302; PubMed Central PMCID: PMC3542458.
388. Onoufriadis A, Paff T, Antony D, Shoemark A, Micha D, Kuyt B, et al. Splice-site mutations in the axonemal outer dynein arm docking complex gene CCDC114 cause primary ciliary dyskinesia. *American journal of human genetics*. 2013;92(1):88-98. Epub 2012/12/25. doi: 10.1016/j.ajhg.2012.11.002. PubMed PMID: 23261303; PubMed Central PMCID: PMC3542455.
389. Wirschell M, Olbrich H, Werner C, Tritschler D, Bower R, Sale WS, et al. The nexin-dynein regulatory complex subunit DRC1 is essential for motile cilia function in algae and humans. *Nature genetics*. 2013;45(3):262-8. Epub 2013/01/29. doi: 10.1038/ng.2533. PubMed PMID: 23354437.
390. Olbrich H, Schmidts M, Werner C, Onoufriadis A, Loges NT, Raidt J, et al. Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *American journal of human genetics*. 2012;91(4):672-84. Epub 2012/10/02. doi: 10.1016/j.ajhg.2012.08.016. PubMed PMID: 23022101; PubMed Central PMCID: PMC3484652.
391. Geremek M, Zietkiewicz E, Diehl SR, Alizadeh BZ, Wijmenga C, Witt M. Linkage analysis localises a Kartagener syndrome gene to a 3.5 cM region on chromosome 15q24-25. *J Med Genet*. 2006;43(1):e1. Epub 2006/01/07. doi: 10.1136/jmg.2005.031526. PubMed PMID: 16397065; PubMed Central PMCID: PMC2564509.
392. Meeks M, Walne A, Spiden S, Simpson H, Mussaffi-Georgy H, Hamam HD, et al. A locus for primary ciliary dyskinesia maps to chromosome 19q. *J Med Genet*. 2000;37(4):241-4. Epub 2000/04/04. PubMed PMID: 10745040; PubMed Central PMCID: PMC1734555.
393. Geremek M, Schoenmaker F, Zietkiewicz E, Pogorzelski A, Diehl S, Wijmenga C, et al. Sequence analysis of 21 genes located in the Kartagener syndrome linkage region on chromosome 15q. *European journal of human genetics : EJHG*. 2008;16(6):688-95. Epub 2008/02/14. doi: 10.1038/ejhg.2008.5. PubMed PMID: 18270537.
394. Narayan D, Krishnan SN, Upender M, Ravikumar TS, Mahoney MJ, Dolan TF, Jr., et al. Unusual inheritance of primary ciliary dyskinesia (Kartagener's

- syndrome). *J Med Genet.* 1994;31(6):493-6. Epub 1994/06/01. PubMed PMID: 8071978; PubMed Central PMCID: PMC1049931.
395. Antony D, Becker-Heck A, Zariwala MA, Schmidts M, Onoufriadis A, Forouhan M, et al. Mutations in CCDC39 and CCDC40 are the major cause of primary ciliary dyskinesia with axonemal disorganization and absent inner dynein arms. *Human mutation.* 2013;34(3):462-72. Epub 2012/12/21. doi: 10.1002/humu.22261. PubMed PMID: 23255504; PubMed Central PMCID: PMC3630464.
396. Loges NT, Olbrich H, Becker-Heck A, Haffner K, Heer A, Reinhard C, et al. Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. *American journal of human genetics.* 2009;85(6):883-9. Epub 2009/12/01. doi: 10.1016/j.ajhg.2009.10.018. PubMed PMID: 19944400; PubMed Central PMCID: PMC2795801.
397. Schwabe GC, Hoffmann K, Loges NT, Birker D, Rossier C, de Santi MM, et al. Primary ciliary dyskinesia associated with normal axoneme ultrastructure is caused by DNAH11 mutations. *Human mutation.* 2008;29(2):289-98. Epub 2007/11/21. doi: 10.1002/humu.20656. PubMed PMID: 18022865.
398. Mazor M, Alkrinawi S, Chalifa-Caspi V, Manor E, Sheffield VC, Aviram M, et al. Primary ciliary dyskinesia caused by homozygous mutation in DNAL1, encoding dynein light chain 1. *American journal of human genetics.* 2011;88(5):599-607. Epub 2011/04/19. doi: 10.1016/j.ajhg.2011.03.018. PubMed PMID: 21496787; PubMed Central PMCID: PMC3146731.
399. Horani A, Druley TE, Zariwala MA, Patel AC, Levinson BT, Van Arendonk LG, et al. Whole-exome capture and sequencing identifies HEATR2 mutation as a cause of primary ciliary dyskinesia. *American journal of human genetics.* 2012;91(4):685-93. Epub 2012/10/09. doi: 10.1016/j.ajhg.2012.08.022. PubMed PMID: 23040496; PubMed Central PMCID: PMC3484505.
400. Budny B, Chen W, Omran H, Fliegauf M, Tzschach A, Wisniewska M, et al. A novel X-linked recessive mental retardation syndrome comprising macrocephaly and ciliary dysfunction is allelic to oral-facial-digital type I syndrome. *Human genetics.* 2006;120(2):171-8. Epub 2006/06/20. doi: 10.1007/s00439-006-0210-5. PubMed PMID: 16783569.
401. Moore A, Escudier E, Roger G, Tamalet A, Pelosse B, Marlin S, et al. RPGR is mutated in patients with a complex X linked phenotype combining primary ciliary dyskinesia and retinitis pigmentosa. *J Med Genet.* 2006;43(4):326-33. Epub 2005/08/02. doi: 10.1136/jmg.2005.034868. PubMed PMID: 16055928; PubMed Central PMCID: PMC2563225.
402. Hjeij R, Lindstrand A, Francis R, Zariwala MA, Liu X, Li Y, et al. ARMC4 Mutations Cause Primary Ciliary Dyskinesia with Randomization of Left/Right Body Asymmetry. *American journal of human genetics.* 2013. Epub 2013/07/16. doi: 10.1016/j.ajhg.2013.06.009. PubMed PMID: 23849778; PubMed Central PMCID: PMC3738828.
403. Tarkar A, Loges NT, Slagle CE, Francis R, Dougherty GW, Tamayo JV, et al. DYX1C1 is required for axonemal dynein assembly and ciliary motility. *Nature genetics.* 2013;45(9):995-1003. Epub 2013/07/23. doi: 10.1038/ng.2707. PubMed PMID: 23872636.

404. Moore DJ, Onoufriadis A, Shoemark A, Simpson MA, Zur Lage PI, de Castro SC, et al. Mutations in ZMYND10, a Gene Essential for Proper Axonemal Assembly of Inner and Outer Dynein Arms in Humans and Flies, Cause Primary Ciliary Dyskinesia. *American journal of human genetics*. 2013. Epub 2013/07/31. doi: 10.1016/j.ajhg.2013.07.009. PubMed PMID: 23891471; PubMed Central PMCID: PMC3738835.
405. Zariwala MA, Gee HY, Kurkowiak M, Al-Mutairi DA, Leigh MW, Hurd TW, et al. ZMYND10 is mutated in primary ciliary dyskinesia and interacts with LRRC6. *American journal of human genetics*. 2013;93(2):336-45. Epub 2013/07/31. doi: 10.1016/j.ajhg.2013.06.007. PubMed PMID: 23891469; PubMed Central PMCID: PMCPMC3738827.
406. Knowles MR, Ostrowski LE, Loges NT, Hurd T, Leigh MW, Huang L, et al. Mutations in SPAG1 cause primary ciliary dyskinesia associated with defective outer and inner dynein arms. *Am J Hum Genet*. 2013;93(4):711-20. Epub 2013/09/24. doi: 10.1016/j.ajhg.2013.07.025. PubMed PMID: 24055112; PubMed Central PMCID: PMCPmc3791252.
407. Austin-Tse C, Halbritter J, Zariwala MA, Gilberti RM, Gee HY, Hellman N, et al. Zebrafish Ciliopathy Screen Plus Human Mutational Analysis Identifies C21orf59 and CCDC65 Defects as Causing Primary Ciliary Dyskinesia. *Am J Hum Genet*. 2013;93(4):672-86. Epub 2013/10/08. doi: 10.1016/j.ajhg.2013.08.015. PubMed PMID: 24094744; PubMed Central PMCID: PMCPmc3791264.
408. Knowles MR, Ostrowski LE, Leigh MW, Sears PR, Davis SD, Wolf WE, et al. Mutations in RSPH1 cause primary ciliary dyskinesia with a unique clinical and ciliary phenotype. *Am J Respir Crit Care Med*. 2014;189(6):707-17. Epub 2014/02/27. doi: 10.1164/rccm.201311-2047OC. PubMed PMID: 24568568; PubMed Central PMCID: PMCPmc3983840.
409. Kurkowiak M, Zietkiewicz E, Witt M. Recent advances in primary ciliary dyskinesia genetics. *J Med Genet*. 2014. Epub 2014/10/30. doi: 10.1136/jmedgenet-2014-102755. PubMed PMID: 25351953.
410. Erzurumluoglu AM, Gaunt TR, Day IN, Baird D, Shihab HA, Richardson TG, et al. Identifying highly-penetrant disease causal mutations using next generation sequencing: Guide to whole process. *BioMed Research International*. 2015;2015(2015). doi: 10.1155/2015/923491.
411. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation*. 2010;2010:baq020. Epub 2010/08/07. doi: 10.1093/database/baq020. PubMed PMID: 20689021; PubMed Central PMCID: PMCPMC2938269.
412. Day IN, Humphries SE, Richards S, Norton D, Reid M. High-throughput genotyping using horizontal polyacrylamide gels with wells arranged for microplate array diagonal gel electrophoresis (MADGE). *Biotechniques*. 1995;19(5):830-5. Epub 1995/11/01. PubMed PMID: 8588924.
413. Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*. 2013;41(Database issue):D43-7. Epub 2012/11/20. doi: 10.1093/nar/gks1068. PubMed PMID: 23161681; PubMed Central PMCID: PMCPMC3531094.

414. Lotta LA, Wu HM, Musallam KM, Peyvandi F. The emerging concept of residual ADAMTS13 activity in ADAMTS13-deficient thrombotic thrombocytopenic purpura. *Blood reviews*. 2013;27(2):71-6. Epub 2013/02/19. doi: 10.1016/j.blre.2013.01.001. PubMed PMID: 23415418.
415. Inacio A, Silva AL, Pinto J, Ji X, Morgado A, Almeida F, et al. Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mRNA decay. *The Journal of biological chemistry*. 2004;279(31):32170-80. Epub 2004/05/27. doi: 10.1074/jbc.M405024200. PubMed PMID: 15161914.
416. Jerber J, Baas D, Soulavie F, Chhin B, Cortier E, Vesque C, et al. The coiled-coil domain containing protein CCDC151 is required for the function of IFT-dependent motile cilia in animals. *Hum Mol Genet*. 2013. Epub 2013/09/27. doi: 10.1093/hmg/ddt445. PubMed PMID: 24067530.
417. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*. 2013;5(3):28. Epub 2013/03/30. doi: 10.1186/gm432. PubMed PMID: 23537139; PubMed Central PMCID: PMC3706896.
418. Vincensini L, Blisnick T, Bastin P. 1001 model organisms to study cilia and flagella. *Biology of the cell / under the auspices of the European Cell Biology Organization*. 2011;103(3):109-30. Epub 2011/02/01. doi: 10.1042/bc20100104. PubMed PMID: 21275904.
419. Schweingruber C, Rufener SC, Zund D, Yamashita A, Muhlemann O. Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochimica et biophysica acta*. 2013;1829(6-7):612-23. Epub 2013/02/26. doi: 10.1016/j.bbagr.2013.02.005. PubMed PMID: 23435113.
420. Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, et al. Functional genomics of the cilium, a sensory organelle. *Current biology : CB*. 2005;15(10):935-41. Epub 2005/05/27. doi: 10.1016/j.cub.2005.04.059. PubMed PMID: 15916950.
421. Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, et al. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell*. 2004;117(4):541-52. Epub 2004/05/13. PubMed PMID: 15137946.
422. Pazour GJ, Agrin N, Leszyk J, Witman GB. Proteomic analysis of a eukaryotic cilium. *The Journal of cell biology*. 2005;170(1):103-13. Epub 2005/07/07. doi: 10.1083/jcb.200504008. PubMed PMID: 15998802; PubMed Central PMCID: PMC171396.
423. Smith JC, Northey JG, Garg J, Pearlman RE, Siu KW. Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of *Tetrahymena thermophila*. *Journal of proteome research*. 2005;4(3):909-19. Epub 2005/06/15. doi: 10.1021/pr050013h. PubMed PMID: 15952738.
424. Stolc V, Samanta MP, Tongprasit W, Marshall WF. Genome-wide transcriptional analysis of flagellar regeneration in *Chlamydomonas reinhardtii* identifies orthologs of ciliary disease genes. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(10):3703-7. Epub 2005/03/02. doi:

- 10.1073/pnas.0408358102. PubMed PMID: 15738400; PubMed Central PMCID: PMCPMC553310.
425. Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T, Subramaniam S, et al. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell*. 2004;117(4):527-39. Epub 2004/05/13. PubMed PMID: 15137945.
426. Rashid S, Breckle R, Hupe M, Geisler S, Doerwald N, Neesen J. The murine *Dnali1* gene encodes a flagellar protein that interacts with the cytoplasmic dynein heavy chain 1. *Molecular Reproduction and Development*. 2006;73(6):784-94. doi: 10.1002/mrd.20475.
427. Zhou J, Yang F, Leu NA, Wang PJ. MNS1 is essential for spermiogenesis and motile ciliary functions in mice. *PLoS genetics*. 2012;8(3):e1002516. Epub 2012/03/08. doi: 10.1371/journal.pgen.1002516. PubMed PMID: 22396656; PubMed Central PMCID: PMCPMC3291534.
428. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, et al. Analysis of *xbx* genes in *C. elegans*. *Development*. 2005;132(8):1923-34. Epub 2005/03/26. doi: 10.1242/dev.01775. PubMed PMID: 15790967.
429. Essner JJ, Amack JD, Nyholm MK, Harris EB, Yost HJ. Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. *Development*. 2005;132(6):1247-60. doi: 10.1242/dev.01663.
430. Dean AB, Mitchell DR. Chlamydomonas ODA10 is a conserved axonemal protein that plays a unique role in outer dynein arm assembly. *Molecular biology of the cell*. 2013;24(23):3689-96. Epub 2013/10/04. doi: 10.1091/mbc.E13-06-0310. PubMed PMID: 24088566; PubMed Central PMCID: PMCPMC3842995.
431. Hjeij R, Onoufriadis A, Watson Christopher M, Slagle Christopher E, Klena Nikolai T, Dougherty Gerard W, et al. CCDC151 Mutations Cause Primary Ciliary Dyskinesia by Disruption of the Outer Dynein Arm Docking Complex Formation. *The American Journal of Human Genetics*. 2014;95(3):257-74. doi: <http://dx.doi.org/10.1016/j.ajhg.2014.08.005>.
432. Alvarez Gonzalez J, Busto Castanon L, Nistal Serrano M. [Evidence for autosomal dominant inheritance through the maternal line in a case of primary ciliary dyskinesia]. *Actas urológicas españolas*. 2006;30(7):728-30. Epub 2006/10/25. PubMed PMID: 17058621.
433. WHO. Intellectual disability Web: World Health Organisation Regional Office for Europe; 2013 [cited 2013 15-11-13]. Available from: <http://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2010/15/childrens-right-to-family-life/definition-intellectual-disability>.
434. Salvador-Carulla L, Rodriguez-Blazquez C, Martorell A. Intellectual disability: an approach from the health sciences perspective. *Salud publica de Mexico*. 2008;50 Suppl 2:s142-50. Epub 2008/05/28. PubMed PMID: 18470341.
435. C. Bessa FL, P. Maciel. Molecular Genetics of Intellectual Disability. Tan PU, editor. Latest Findings in Intellectual and Developmental Disabilities Research: InTech; 2012.

436. Centers for Disease Control and Prevention C. Economic costs associated with mental retardation, cerebral palsy, hearing loss, and vision impairment--United States, 2003. *MMWR Morbidity and mortality weekly report*. 2004;53(3):57-9. Epub 2004/01/30. PubMed PMID: 14749614.
437. Leonard H, Wen X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. *Mental retardation and developmental disabilities research reviews*. 2002;8(3):117-34. Epub 2002/09/07. doi: 10.1002/mrdd.10031. PubMed PMID: 12216056.
438. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *American journal of medical genetics Part A*. 2006;140(19):2063-74. Epub 2006/08/19. doi: 10.1002/ajmg.a.31416. PubMed PMID: 16917849.
439. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. *Nature genetics*. 2011;43(9):838-46. Epub 2011/08/16. doi: 10.1038/ng.909. PubMed PMID: 21841781; PubMed Central PMCID: PMC3171215.
440. Presson AP, Partyka G, Jensen KM, Devine OJ, Rasmussen SA, McCabe LL, et al. Current estimate of down syndrome population prevalence in the United States. *J Pediatr*. 2013;163(4):1163-8. Epub 2013/07/28. doi: 10.1016/j.jpeds.2013.06.013. PubMed PMID: 23885965.
441. Morrow EM. Genomic copy number variation in disorders of cognitive development. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2010;49(11):1091-104. Epub 2010/10/26. doi: 10.1016/j.jaac.2010.08.009. PubMed PMID: 20970697; PubMed Central PMCID: PMC3137887.
442. Hagerman PJ, Hagerman RJ. The fragile-X premutation: a maturing perspective. *American journal of human genetics*. 2004;74(5):805-16. Epub 2004/03/31. doi: 10.1086/386296. PubMed PMID: 15052536; PubMed Central PMCID: PMC31181976.
443. Coffee B, Keith K, Albizua I, Malone T, Mowrey J, Sherman SL, et al. Incidence of fragile X syndrome by newborn screening for methylated FMR1 DNA. *American journal of human genetics*. 2009;85(4):503-14. Epub 2009/10/07. doi: 10.1016/j.ajhg.2009.09.007. PubMed PMID: 19804849; PubMed Central PMCID: PMC312756550.
444. Lubs HA, Stevenson RE, Schwartz CE. Fragile X and X-linked intellectual disability: four decades of discovery. *American journal of human genetics*. 2012;90(4):579-90. Epub 2012/04/10. doi: 10.1016/j.ajhg.2012.02.018. PubMed PMID: 22482801; PubMed Central PMCID: PMC3322227.
445. Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V, et al. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Archives of general psychiatry*. 2009;66(9):947-56. Epub 2009/09/09. doi: 10.1001/archgenpsychiatry.2009.80. PubMed PMID: 19736351; PubMed Central PMCID: PMC312958844.
446. Alazami AM, Hijazi H, Al-Dosari MS, Shaheen R, Hashem A, Aldahmesh MA, et al. Mutation in ADAT3, encoding adenosine deaminase acting on transfer

- RNA, causes intellectual disability and strabismus. *J Med Genet*. 2013;50(7):425-30. Epub 2013/04/27. doi: 10.1136/jmedgenet-2012-101378. PubMed PMID: 23620220.
447. Hart TC, Hart PS, Bowden DW, Michalec MD, Callison SA, Walker SJ, et al. Mutations of the cathepsin C gene are responsible for Papillon-Lefevre syndrome. *J Med Genet*. 1999;36(12):881-7. Epub 1999/12/14. PubMed PMID: 10593994; PubMed Central PMCID: PMC1734286.
448. Nagy N, Valyi P, Csoma Z, Sulak A, Tripolszki K, Farkas K, et al. CTSC and Papillon-Lefevre syndrome: detection of recurrent mutations in Hungarian patients, a review of published variants and database update. *Molecular genetics & genomic medicine*. 2014;2(3):217-28. Epub 2014/06/18. doi: 10.1002/mgg3.61. PubMed PMID: 24936511; PubMed Central PMCID: PMC1734286.
449. Turk D, Janjic V, Stern I, Podobnik M, Lamba D, Dahl SW, et al. Structure of human dipeptidyl peptidase I (cathepsin C): exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases. *The EMBO journal*. 2001;20(23):6570-82. Epub 2001/12/01. doi: 10.1093/emboj/20.23.6570. PubMed PMID: 11726493; PubMed Central PMCID: PMC125750.
450. ESP NG. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP) Web2013 [cited 2013 Dec]. Available from: <http://evs.gs.washington.edu/EVS/>.
451. Gaunt TR, Cooper JA, Miller GJ, Day IN, O'Dell SD. Positive associations between single nucleotide polymorphisms in the IGF2 gene region and body mass index in adult males. *Hum Mol Genet*. 2001;10(14):1491-501. Epub 2001/07/13. PubMed PMID: 11448941.
452. Zhang Y, Lundgren T, Renvert S, Tatakis DN, Firatli E, Uygur C, et al. Evidence of a founder effect for four cathepsin C gene mutations in Papillon-Lefevre syndrome patients. *J Med Genet*. 2001;38(2):96-101. Epub 2001/02/07. PubMed PMID: 11158173; PubMed Central PMCID: PMC1734811.
453. Yu JH, Harrell TM, Jamal SM, Tabor HK, Bamshad MJ. Attitudes of genetics professionals toward the return of incidental results from exome and whole-genome sequencing. *American journal of human genetics*. 2014;95(1):77-84. Epub 2014/07/01. doi: 10.1016/j.ajhg.2014.06.004. PubMed PMID: 24975944; PubMed Central PMCID: PMC1734811.
454. Rabbani B, Tekin M, Mahdiah N. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*. 2014;59(1):5-15. Epub 2013/11/08. doi: 10.1038/jhg.2013.114. PubMed PMID: 24196381.
455. Uitto J, Christiano AM, McLean WH, McGrath JA. Novel molecular therapies for heritable skin disorders. *The Journal of investigative dermatology*. 2012;132(3 Pt 2):820-8. Epub 2011/12/14. doi: 10.1038/jid.2011.389. PubMed PMID: 22158553; PubMed Central PMCID: PMC1734811.
456. Takeichi T, Nanda A, Aristodemou S, McMillan JR, Lee J, Akiyama M, et al. Whole-exome sequencing diagnosis of two autosomal recessive disorders in one family. *The British journal of dermatology*. 2014. Epub 2014/10/14. doi: 10.1111/bjd.13473. PubMed PMID: 25308318.
457. Erzurumluoglu AM, Shihab HA, Rodriguez S, Gaunt TR, Day INM. Importance of genetic studies in consanguineous populations to characterization of human gene function. *Annals of human genetics*. 2016.

458. Mayo O. A century of Hardy-Weinberg equilibrium. *Twin research and human genetics : the official journal of the International Society for Twin Studies*. 2008;11(3):249-56. Epub 2008/05/24. doi: 10.1375/twin.11.3.249. PubMed PMID: 18498203.
459. Erzurumluoglu AM, Alsaadi MM, Rodriguez S, Alotaibi TS, Guthrie PA, Lewis S, et al. Proxy molecular diagnosis from whole-exome sequencing reveals Papillon-Lefevre syndrome caused by a missense mutation in CTSC. *PloS one*. 2015;10(3):e0121351. doi: 10.1371/journal.pone.0121351. PubMed PMID: 25799584; PubMed Central PMCID: PMC4370501.
460. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International journal of epidemiology*. 2013;42(1):97-110. Epub 2012/04/18. doi: 10.1093/ije/dys066. PubMed PMID: 22507742; PubMed Central PMCID: PMC3600619.
461. Miklos GL, Yamamoto M, Burns RG, Maleszka R. An essential cell division gene of *Drosophila*, absent from *Saccharomyces*, encodes an unusual protein with tubulin-like and myosin-like peptide motifs. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94(10):5189-94. Epub 1997/05/13. PubMed PMID: 9144213; PubMed Central PMCID: PMCPMC24654.
462. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. *Nature reviews Genetics*. 2011;12(10):692-702. Epub 2011/09/01. doi: 10.1038/nrg3053. PubMed PMID: 21878963.
463. Watkins S, Madison J, Galliano M, Minchiotti L, Putnam FW. Analbuminemia: three cases resulting from different point mutations in the albumin gene. *Proceedings of the National Academy of Sciences of the United States of America*. 1994;91(20):9417-21. Epub 1994/09/27. PubMed PMID: 7937781; PubMed Central PMCID: PMCPMC44823.
464. Farrugia A. Albumin usage in clinical medicine: tradition or therapeutic? *Transfusion medicine reviews*. 2010;24(1):53-63. Epub 2009/12/08. doi: 10.1016/j.tmr.2009.09.005. PubMed PMID: 19962575.
465. Koot BG, Houwen R, Pot DJ, Nauta J. Congenital analbuminaemia: biochemical and clinical implications. A case report and literature review. *European journal of pediatrics*. 2004;163(11):664-70. Epub 2004/08/10. doi: 10.1007/s00431-004-1492-z. PubMed PMID: 15300429.
466. Slatis HM. A method of estimating the frequency of abnormal autosomal recessive genes in man. *American journal of human genetics*. 1954;6(4):412-8. Epub 1954/12/01. PubMed PMID: 14349946; PubMed Central PMCID: PMCPMC1716582.
467. Morton NE, Crow JF, Muller HJ. AN ESTIMATE OF THE MUTATIONAL DAMAGE IN MAN FROM DATA ON CONSANGUINEOUS MARRIAGES. *Proceedings of the National Academy of Sciences of the United States of America*. 1956;42(11):855-63. Epub 1956/11/01. PubMed PMID: 16589958; PubMed Central PMCID: PMCPMC528351.
468. Bhuiyan ZA, Momenah TS, Gong Q, Amin AS, Ghamdi SA, Carvalho JS, et al. Recurrent intrauterine fetal loss due to near absence of HERG: clinical and functional characterization of a homozygous nonsense HERG Q1070X mutation. *Heart rhythm : the official journal of the Heart Rhythm Society*. 2008;5(4):553-61. Epub 2008/03/26.

- doi: 10.1016/j.hrthm.2008.01.020. PubMed PMID: 18362022; PubMed Central PMCID: PMC2682734.
469. Saad FA, Jauniaux E. Recurrent early pregnancy loss and consanguinity. *Reproductive biomedicine online*. 2002;5(2):167-70. Epub 2002/11/07. PubMed PMID: 12419042.
470. Jaber L, Merlob P, Gabriel R, Shohat M. Effects of consanguineous marriage on reproductive outcome in an Arab community in Israel. *J Med Genet*. 1997;34(12):1000-2. Epub 1998/01/16. PubMed PMID: 9429142; PubMed Central PMCID: PMC2682734.
471. Satoh JI, Tokumoto H, Kurohara K, Yukitake M, Matsui M, Kuroda Y, et al. Adult-onset Krabbe disease with homozygous T1853C mutation in the galactocerebrosidase gene. Unusual MRI findings of corticospinal tract demyelination. *Neurology*. 1997;49(5):1392-9. Epub 1997/12/31. PubMed PMID: 9371928.
472. Sobel E, Davanipour Z, Alter M. Genetic analysis of late-onset diseases using first-degree relatives. *Neuroepidemiology*. 1988;7(2):81-8. Epub 1988/01/01. PubMed PMID: 3374730.
473. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. The DNA sequence of the human X chromosome. *Nature*. 2005;434(7031):325-37. Epub 2005/03/18. doi: 10.1038/nature03440. PubMed PMID: 15772651; PubMed Central PMCID: PMC2665286.
474. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature genetics*. 2009;41(5):535-43. Epub 2009/04/21. doi: 10.1038/ng.367. PubMed PMID: 19377476; PubMed Central PMCID: PMC2872007.
475. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, et al. Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature*. 1997;387(6636):903-8. Epub 1997/06/26. doi: 10.1038/43185. PubMed PMID: 9202122.
476. Kingsmore S. Comprehensive carrier screening and molecular diagnostic testing for recessive childhood diseases. *PLoS currents*. 2012:e4f9877ab8ffa9. Epub 2012/08/09. doi: 10.1371/4f9877ab8ffa9. PubMed PMID: 22872815; PubMed Central PMCID: PMC3392137.
477. Raal FJ, Santos RD. Homozygous familial hypercholesterolemia: current perspectives on diagnosis and treatment. *Atherosclerosis*. 2012;223(2):262-8. Epub 2012/03/09. doi: 10.1016/j.atherosclerosis.2012.02.019. PubMed PMID: 22398274.
478. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet*. 2010;19:R125-R30.
479. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human mutation*. 2013;34(9):E2393-402. Epub 2013/07/12. doi: 10.1002/humu.22376. PubMed PMID: 23843252; PubMed Central PMCID: PMC4109890.
480. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. Genetic markers and population history: Finland revisited. *European journal of human genetics : EJHG*.

- 2009;17(10):1336-46. Epub 2009/04/16. doi: 10.1038/ejhg.2009.53. PubMed PMID: 19367325; PubMed Central PMCID: PMCPMC2986642.
481. Sillanpaa MJ. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* (Edinb). 2011;106(4):511-9. Epub 2010/07/16. doi: 10.1038/hdy.2010.91. PubMed PMID: 20628415; PubMed Central PMCID: PMCPMC3183892.
482. Kim TJ. Riyadh. In: Britannica EoTE, editor. *Encyclopædia Britannica*. Encyclopædia Britannica Online: Encyclopædia Britannica Inc; 2013.
483. El-Mouzan MI, Al-Salloum AA, Al-Herbish AS, Qurachi MM, Al-Omar AA. Regional variations in the prevalence of consanguinity in Saudi Arabia. *Saudi medical journal*. 2007;28(12):1881-4. Epub 2007/12/07. PubMed PMID: 18060221.
484. Alkuraya FS. Genetics and genomic medicine in Saudi Arabia. *Molecular genetics & genomic medicine*. 2014;2(5):369-78. doi: 10.1002/mgg3.97.
485. R.V.R. Chandrasekhara Rao SVW. Andhra Pradesh. In: Britannica EoTE, editor. *Encyclopædia Britannica*. Web: Encyclopædia Britannica Online; 2013.
486. Kumar D. *Genetic Disorders of the Indian Subcontinent*: Springer; 2004.
487. Nelson MJ. Pakistan in 2008: Moving beyond Musharraf. *Asian Survey*. 2009;49(1):16-27. doi: 10.1525/as.2009.49.1.16.
488. Sheridan E, Wright J, Small N, Corry PC, Oddie S, Whibley C, et al. Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *Lancet*. 2013;382(9901):1350-9. Epub 2013/07/09. doi: 10.1016/s0140-6736(13)61132-0. PubMed PMID: 23830354.
489. Islam MM. The practice of consanguineous marriage in Oman: prevalence, trends and determinants. *Journal of biosocial science*. 2012;44(5):571-94. Epub 2012/02/10. doi: 10.1017/s0021932012000016. PubMed PMID: 22317781.
490. Saha N, el Sheikh FS. Inbreeding levels in Khartoum. *Journal of biosocial science*. 1988;20(3):333-6. Epub 1988/07/01. PubMed PMID: 3215913.
491. Sueyoshi S, Ohtsuka R. Effects of polygyny and consanguinity on high fertility in the rural Arab population in South Jordan. *Journal of biosocial science*. 2003;35(4):513-26. Epub 2003/11/19. PubMed PMID: 14621249.
492. Boston RC, Sumner AE. STATA: a statistical analysis system for examining biomedical data. *Advances in experimental medicine and biology*. 2003;537:353-69. Epub 2004/03/05. PubMed PMID: 14995047.
493. Vega GL, Grundy SM. Mechanisms of primary hypercholesterolemia in humans. *American heart journal*. 1987;113(2 Pt 2):493-502. Epub 1987/02/01. PubMed PMID: 3544763.
494. Roche SL, Silversides CK. Hypertension, obesity, and coronary artery disease in the survivors of congenital heart disease. *The Canadian journal of cardiology*. 2013;29(7):841-8. Epub 2013/05/22. doi: 10.1016/j.cjca.2013.03.021. PubMed PMID: 23688771.
495. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS medicine*. 2007;4(12):e352. Epub 2007/12/14. doi: 10.1371/journal.pmed.0040352. PubMed PMID: 18076282; PubMed Central PMCID: PMCPMC2121108.

496. Khajavi M, Inoue K, Lupski JR. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *European journal of human genetics : EJHG*. 2006;14(10):1074-81. Epub 2006/06/08. doi: 10.1038/sj.ejhg.5201649. PubMed PMID: 16757948.
497. Khalak HG, Wakil SM, Imtiaz F, Ramzan K, Baz B, Almostafa A, et al. Autozygome maps dispensable DNA and reveals potential selective bias against nullizygosity. *Genet Med*. 2012;14(5):515-9. Epub 2012/01/14. doi: 10.1038/gim.2011.28. PubMed PMID: 22241088.
498. Alsalem AB, Halees AS, Anazi S, Alshamekh S, Alkuraya FS. Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. *PLoS genetics*. 2013;9(12):e1004030. Epub 2013/12/25. doi: 10.1371/journal.pgen.1004030. PubMed PMID: 24367280; PubMed Central PMCID: PMC3868571.
499. Narasimhan V, Hunt K, Mason D, Baker CL, Karczewski K, Barnes M, et al. Health and population effects of rare gene knockouts in adult humans with related parents. *bioRxiv*. 2015. doi: 10.1101/031641.
500. Saleheen D, Natarajan P, Zhao W, Rasheed A, Khetarpal S, Won H-H, et al. Human knockouts in a cohort with a high rate of consanguinity. *bioRxiv*. 2015. doi: 10.1101/031518.
501. Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, et al. Directional dominance on stature and cognition in diverse human populations. *Nature*. 2015;523(7561):459-62. Epub 2015/07/02. doi: 10.1038/nature14618. PubMed PMID: 26131930; PubMed Central PMCID: PMC4516141.
502. Dyer O. MP is criticised for saying that marriage of first cousins is a health problem. *BMJ*. 2005;331(7528):1292. doi: 10.1136/bmj.331.7528.1292.
503. Hamamy H. Consanguineous marriages : Preconception consultation in primary health care settings. *Journal of community genetics*. 2012;3(3):185-92. Epub 2011/11/24. doi: 10.1007/s12687-011-0072-y. PubMed PMID: 22109912; PubMed Central PMCID: PMC3419292.
504. Erzurumluoglu M. Consanguineous Marriages: Perspectives from Social Taboos, Religion and Science. *The Fountain*. 2014;99(May-June 2014).
505. Lenay C. Hugo De Vries: from the theory of intracellular pangenesis to the rediscovery of Mendel. *Comptes rendus de l'Academie des sciences Serie III, Sciences de la vie*. 2000;323(12):1053-60. Epub 2001/01/09. PubMed PMID: 11147091.
506. Gregory SG, Barlow KF, McLay KE, Kaul R, Swarbreck D, Dunham A, et al. The DNA sequence and biological annotation of human chromosome[thinsp]1. *Nature*. 2006;441(7091):315-21. doi: http://www.nature.com/nature/journal/v441/n7091/supinfo/nature04727_S1.html.
507. Casey JP, McGettigan PA, Healy F, Hogg C, Reynolds A, Kennedy BN, et al. Unexpected genetic heterogeneity for primary ciliary dyskinesia in the Irish Traveller population. *Eur J Hum Genet*. 2014. Epub 2014/05/16. doi: 10.1038/ejhg.2014.79. PubMed PMID: 24824133.
508. Burgoyne T, Lewis A, Dewar A, Luther P, Hogg C, Shoemark A, et al. Characterizing the ultrastructure of primary ciliary dyskinesia transposition defect

- using electron tomography. *Cytoskeleton* (Hoboken). 2014;71(5):294-301. Epub 2014/03/13. doi: 10.1002/cm.21171. PubMed PMID: 24616277.
509. Zietkiewicz E, Bukowy-Bieryllo Z, Voelkel K, Klimek B, Dmenska H, Pogorzelski A, et al. Mutations in radial spoke head genes and ultrastructural cilia defects in East-European cohort of primary ciliary dyskinesia patients. *PLoS One*. 2012;7(3):e33667. Epub 2012/03/27. doi: 10.1371/journal.pone.0033667. PubMed PMID: 22448264; PubMed Central PMCID: PMC3308995.
510. Onoufriadis A, Shoemark A, Schmidts M, Patel M, Jimenez G, Liu H, et al. Targeted NGS gene panel identifies mutations in RSPH1 causing primary ciliary dyskinesia and a common mechanism for ciliary central pair agenesis due to radial spoke defects. *Hum Mol Genet*. 2014;23(13):3362-74. Epub 2014/02/13. doi: 10.1093/hmg/ddu046. PubMed PMID: 24518672; PubMed Central PMCID: PMC34049301.
511. Fliegauf M, Olbrich H, Horvath J, Wildhaber JH, Zariwala MA, Kennedy M, et al. Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia. *Am J Respir Crit Care Med*. 2005;171(12):1343-9. Epub 2005/03/08. doi: 10.1164/rccm.200411-1583OC. PubMed PMID: 15750039; PubMed Central PMCID: PMC14718478.
512. Tate G, Tajiri T, Kishimoto K, Mitsuya T. A novel mutation of the axonemal dynein heavy chain gene 5 (DNAH5) in a Japanese neonate with asplenia syndrome. *Med Mol Morphol*. 2014. Epub 2014/06/11. doi: 10.1007/s00795-014-0079-7. PubMed PMID: 24912412.
513. Jerber J, Baas D, Soulavie F, Chhin B, Cortier E, Vesque C, et al. The coiled-coil domain containing protein CCDC151 is required for the function of IFT-dependent motile cilia in animals. *Hum Mol Genet*. 2014;23(3):563-77. Epub 2013/09/27. doi: 10.1093/hmg/ddt445. PubMed PMID: 24067530.
514. Teves ME, Zhang Z, Costanzo RM, Henderson SC, Corwin FD, Zweit J, et al. Sperm-associated antigen-17 gene is essential for motile cilia function and neonatal survival. *Am J Respir Cell Mol Biol*. 2013;48(6):765-72. Epub 2013/02/19. doi: 10.1165/rcmb.2012-0362OC. PubMed PMID: 23418344; PubMed Central PMCID: PMC3727877.
515. Geremek M, Zietkiewicz E, Bruinenberg M, Franke L, Pogorzelski A, Wijmenga C, et al. Ciliary genes are down-regulated in bronchial tissue of primary ciliary dyskinesia patients. *PLoS One*. 2014;9(2):e88216. Epub 2014/02/12. doi: 10.1371/journal.pone.0088216. PubMed PMID: 24516614; PubMed Central PMCID: PMC3916409.
516. Bartoloni L, Blouin JL, Pan Y, Gehrig C, Maiti AK, Scamuffa N, et al. Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of situs inversus totalis and most likely primary ciliary dyskinesia. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(16):10282-6. Epub 2002/07/27. doi: 10.1073/pnas.152337699. PubMed PMID: 12142464; PubMed Central PMCID: PMC124905.
517. Olbrich H, Schmidts M, Werner C, Onoufriadis A, Loges NT, Raidt J, et al. Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *Am J Hum Genet*. 2012;91(4):672-84.

- Epub 2012/10/02. doi: 10.1016/j.ajhg.2012.08.016. PubMed PMID: 23022101; PubMed Central PMCID: PMC3484652.
518. Onoufriadis A, Paff T, Antony D, Shoemark A, Micha D, Kuyt B, et al. Splice-site mutations in the axonemal outer dynein arm docking complex gene *CCDC114* cause primary ciliary dyskinesia. *Am J Hum Genet.* 2013;92(1):88-98. Epub 2012/12/25. doi: 10.1016/j.ajhg.2012.11.002. PubMed PMID: 23261303; PubMed Central PMCID: PMC3542455.
519. Horani A, Brody SL, Ferkol TW, Shoseyov D, Wasserman MG, Ta-Shma A, et al. *CCDC65* Mutation Causes Primary Ciliary Dyskinesia with Normal Ultrastructure and Hyperkinetic Cilia. *PloS one.* 2013;8(8):e72299. Epub 2013/08/31. doi: 10.1371/journal.pone.0072299. PubMed PMID: 23991085; PubMed Central PMCID: PMC3753302.
520. Moore DJ, Onoufriadis A, Shoemark A, Simpson MA, zur Lage PI, de Castro SC, et al. Mutations in *ZMYND10*, a gene essential for proper axonemal assembly of inner and outer dynein arms in humans and flies, cause primary ciliary dyskinesia. *Am J Hum Genet.* 2013;93(2):346-56. Epub 2013/07/31. doi: 10.1016/j.ajhg.2013.07.009. PubMed PMID: 23891471; PubMed Central PMCID: PMC3738835.
521. Onoufriadis A, Shoemark A, Munye MM, James CT, Schmidts M, Patel M, et al. Combined exome and whole-genome sequencing identifies mutations in *ARMC4* as a cause of primary ciliary dyskinesia with defects in the outer dynein arm. *J Med Genet.* 2014;51(1):61-7. Epub 2013/11/10. doi: 10.1136/jmedgenet-2013-101938. PubMed PMID: 24203976; PubMed Central PMCID: PMC3888613.
522. Tarkar A, Loges NT, Slagle CE, Francis R, Dougherty GW, Tamayo JV, et al. *DYX1C1* is required for axonemal dynein assembly and ciliary motility. *Nat Genet.* 2013;45(9):995-1003. Epub 2013/07/23. doi: 10.1038/ng.2707. PubMed PMID: 23872636; PubMed Central PMCID: PMC3400444.
523. Ben Khelifa M, Coutton C, Zouari R, Karaouzene T, Rendu J, Bidart M, et al. Mutations in *DNAH1*, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *Am J Hum Genet.* 2014;94(1):95-104. Epub 2013/12/24. doi: 10.1016/j.ajhg.2013.11.017. PubMed PMID: 24360805; PubMed Central PMCID: PMC3882734.
524. Escudier E, Duquesnoy P, Papon JF, Amselem S. Ciliary defects and genetics of primary ciliary dyskinesia. *Paediatr Respir Rev.* 2009;10(2):51-4. Epub 2009/05/05. doi: 10.1016/j.prrv.2009.02.001. PubMed PMID: 19410201.
525. Zariwala M, O'Neal WK, Noone PG, Leigh MW, Knowles MR, Ostrowski LE. Investigation of the possible role of a novel gene, *DPCD*, in primary ciliary dyskinesia. *Am J Respir Cell Mol Biol.* 2004;30(4):428-34. Epub 2003/11/25. doi: 10.1165/rcmb.2003-0338RC. PubMed PMID: 14630615.
526. Molinari F, Rio M, Meskenaitė V, Encha-Razavi F, Auge J, Bacq D, et al. Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation. *Science (New York, NY).* 2002;298(5599):1779-81. Epub 2002/12/03. doi: 10.1126/science.1076521. PubMed PMID: 12459588.
527. Higgins JJ, Pucilowska J, Lombardi RQ, Rooney JP. A mutation in a novel ATP-dependent Lon protease gene in a kindred with mild mental retardation.

- Neurology. 2004;63(10):1927-31. Epub 2004/11/24. PubMed PMID: 15557513; PubMed Central PMCID: PMC1201536.
528. Basel-Vanagaite L, Attia R, Yahav M, Ferland RJ, Anteki L, Walsh CA, et al. The CC2D1A, a member of a new gene family with C2 domains, is involved in autosomal recessive non-syndromic mental retardation. *Journal of medical genetics*. 2006;43(3):203-10. Epub 2005/07/22. doi: 10.1136/jmg.2005.035709. PubMed PMID: 16033914; PubMed Central PMCID: PMC12563235.
529. Motazacker MM, Rost BR, Hucho T, Garshasbi M, Kahrizi K, Ullmann R, et al. A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. *American journal of human genetics*. 2007;81(4):792-8. Epub 2007/09/12. doi: 10.1086/521275. PubMed PMID: 17847003; PubMed Central PMCID: PMC1227928.
530. Garshasbi M, Hadavi V, Habibi H, Kahrizi K, Kariminejad R, Behjati F, et al. A defect in the TUSC3 gene is associated with autosomal recessive mental retardation. *American journal of human genetics*. 2008;82(5):1158-64. Epub 2008/05/03. doi: 10.1016/j.ajhg.2008.03.018. PubMed PMID: 18452889; PubMed Central PMCID: PMC12651624.
531. Molinari F, Foulquier F, Tarpey PS, Morelle W, Boissel S, Teague J, et al. Oligosaccharyltransferase-subunit mutations in nonsyndromic mental retardation. *American journal of human genetics*. 2008;82(5):1150-7. Epub 2008/05/06. doi: 10.1016/j.ajhg.2008.03.021. PubMed PMID: 18455129; PubMed Central PMCID: PMC12427205.
532. Khan MA, Rafiq MA, Noor A, Ali N, Ali G, Vincent JB, et al. A novel deletion mutation in the TUSC3 gene in a consanguineous Pakistani family with autosomal recessive nonsyndromic intellectual disability. *BMC medical genetics*. 2011;12:56. Epub 2011/04/26. doi: 10.1186/1471-2350-12-56. PubMed PMID: 21513506; PubMed Central PMCID: PMC13096909.
533. Mir A, Kaufman L, Noor A, Motazacker MM, Jamil T, Azam M, et al. Identification of mutations in TRAPPC9, which encodes the NIK- and IKK-beta-binding protein, in nonsyndromic autosomal-recessive mental retardation. *American journal of human genetics*. 2009;85(6):909-15. Epub 2009/12/17. doi: 10.1016/j.ajhg.2009.11.009. PubMed PMID: 20004765; PubMed Central PMCID: PMC12790571.
534. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*. 2011;478(7367):57-63. Epub 2011/09/23. doi: 10.1038/nature10423. PubMed PMID: 21937992.
535. Philippe O, Rio M, Carioux A, Plaza JM, Guigue P, Molinari F, et al. Combination of linkage mapping and microarray-expression analysis identifies NF-kappaB signaling defect as a cause of autosomal-recessive mental retardation. *American journal of human genetics*. 2009;85(6):903-8. Epub 2009/12/17. doi: 10.1016/j.ajhg.2009.11.007. PubMed PMID: 20004764; PubMed Central PMCID: PMC12795800.
536. Abou Jamra R, Wohlfart S, Zweier M, Uebe S, Priebe L, Ekici A, et al. Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. *European journal*

- of human genetics : EJHG. 2011;19(11):1161-6. Epub 2011/06/02. doi: 10.1038/ejhg.2011.98. PubMed PMID: 21629298; PubMed Central PMCID: PMC3198153.
537. Mochida GH, Mahajnah M, Hill AD, Basel-Vanagaite L, Gleason D, Hill RS, et al. A truncating mutation of TRAPPC9 is associated with autosomal-recessive intellectual disability and postnatal microcephaly. *American journal of human genetics*. 2009;85(6):897-902. Epub 2009/12/17. doi: 10.1016/j.ajhg.2009.10.027. PubMed PMID: 20004763; PubMed Central PMCID: PMC3198153.
538. Kakar N, Goebel I, Daud S, Nurnberg G, Agha N, Ahmad A, et al. A homozygous splice site mutation in TRAPPC9 causes intellectual disability and microcephaly. *European journal of medical genetics*. 2012;55(12):727-31. Epub 2012/09/20. doi: 10.1016/j.ejmg.2012.08.010. PubMed PMID: 22989526.
539. Marangi G, Leuzzi V, Manti F, Lattante S, Orteschi D, Pecile V, et al. TRAPPC9-related autosomal recessive intellectual disability: report of a new mutation and clinical phenotype. *European journal of human genetics : EJHG*. 2013;21(2):229-32. Epub 2012/05/03. doi: 10.1038/ejhg.2012.79. PubMed PMID: 22549410; PubMed Central PMCID: PMC3548258.
540. Pak C, Garshasbi M, Kahrizi K, Gross C, Apponi LH, Noto JJ, et al. Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in Drosophila and humans. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(30):12390-5. Epub 2011/07/08. doi: 10.1073/pnas.1107103108. PubMed PMID: 21734151; PubMed Central PMCID: PMC3145741.
541. Hashimoto S, Boissel S, Zarhrate M, Rio M, Munnich A, Egly JM, et al. MED23 mutation links intellectual disability to dysregulation of immediate early gene expression. *Science (New York, NY)*. 2011;333(6046):1161-3. Epub 2011/08/27. doi: 10.1126/science.1206638. PubMed PMID: 21868677.
542. Caliskan M, Chong JX, Uricchio L, Anderson R, Chen P, Sougnez C, et al. Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECP gene on chromosome 19p13. *Human molecular genetics*. 2011;20(7):1285-9. Epub 2011/01/08. doi: 10.1093/hmg/ddq569. PubMed PMID: 21212097; PubMed Central PMCID: PMC3115579.
543. Rafiq MA, Kuss AW, Puettmann L, Noor A, Ramiah A, Ali G, et al. Mutations in the alpha 1,2-mannosidase gene, MAN1B1, cause autosomal-recessive intellectual disability. *American journal of human genetics*. 2011;89(1):176-82. Epub 2011/07/19. doi: 10.1016/j.ajhg.2011.06.006. PubMed PMID: 21763484; PubMed Central PMCID: PMC3135808.
544. Alazami AM, Al-Owain M, Alzahrani F, Shuaib T, Al-Shamrani H, Al-Falki YH, et al. Loss of function mutation in LARP7, chaperone of 7SK ncRNA, causes a syndrome of facial dysmorphism, intellectual disability, and primordial dwarfism. *Human mutation*. 2012;33(10):1429-34. Epub 2012/08/07. doi: 10.1002/humu.22175. PubMed PMID: 22865833.
545. Hu H, Eggers K, Chen W, Garshasbi M, Motazacker MM, Wroegemann K, et al. ST3GAL3 mutations impair the development of higher cognitive functions. *American journal of human genetics*. 2011;89(3):407-14. Epub 2011/09/13. doi:

- 10.1016/j.ajhg.2011.08.008. PubMed PMID: 21907012; PubMed Central PMCID: PMCpMc3169827.
546. Abbasi-Moheb L, Mertel S, Gonsior M, Nouri-Vahid L, Kahrizi K, Cirak S, et al. Mutations in NSUN2 cause autosomal-recessive intellectual disability. *American journal of human genetics*. 2012;90(5):847-55. Epub 2012/05/01. doi: 10.1016/j.ajhg.2012.03.021. PubMed PMID: 22541559; PubMed Central PMCID: PMCpMc3376487.
547. Khan MA, Rafiq MA, Noor A, Hussain S, Flores JV, Rupp V, et al. Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *American journal of human genetics*. 2012;90(5):856-63. Epub 2012/05/01. doi: 10.1016/j.ajhg.2012.03.023. PubMed PMID: 22541562; PubMed Central PMCID: PMCpMc3376419.
548. Noor A, Windpassinger C, Patel M, Stachowiak B, Mikhailov A, Azam M, et al. CC2D2A, encoding a coiled-coil and C2 domain protein, causes autosomal-recessive mental retardation with retinitis pigmentosa. *American journal of human genetics*. 2008;82(4):1011-8. Epub 2008/04/05. doi: 10.1016/j.ajhg.2008.01.021. PubMed PMID: 18387594; PubMed Central PMCID: PMCpMc2427291.
549. Corbett MA, Bahlo M, Jolly L, Afawi Z, Gardner AE, Oliver KL, et al. A focal epilepsy and intellectual disability syndrome is due to a mutation in TBC1D24. *American journal of human genetics*. 2010;87(3):371-5. Epub 2010/08/28. doi: 10.1016/j.ajhg.2010.08.001. PubMed PMID: 20797691; PubMed Central PMCID: PMCpMc2933342.
550. Koehler K, Malik M, Mahmood S, Giesselmann S, Beetz C, Hennings JC, et al. Mutations in GMPPA cause a glycosylation disorder characterized by intellectual disability and autonomic dysfunction. *American journal of human genetics*. 2013;93(4):727-34. Epub 2013/09/17. doi: 10.1016/j.ajhg.2013.08.002. PubMed PMID: 24035193; PubMed Central PMCID: PMCpMc3791256.
551. Kvarnung M, Nilsson D, Lindstrand A, Korenke GC, Chiang SC, Blennow E, et al. A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. *Journal of medical genetics*. 2013;50(8):521-8. Epub 2013/05/03. doi: 10.1136/jmedgenet-2013-101654. PubMed PMID: 23636107.
552. Kong XF, Bousfiha A, Rouissi A, Itan Y, Abhyankar A, Bryant V, et al. A novel homozygous p.R1105X mutation of the AP4E1 gene in twins with hereditary spastic paraplegia and mycobacterial disease. *PloS one*. 2013;8(3):e58286. Epub 2013/03/09. doi: 10.1371/journal.pone.0058286. PubMed PMID: 23472171; PubMed Central PMCID: PMCpMc3589270.
553. Memon MM, Raza SI, Basit S, Kousar R, Ahmad W, Ansar M. A novel WDR62 mutation causes primary microcephaly in a Pakistani family. *Molecular biology reports*. 2013;40(1):591-5. Epub 2012/10/16. doi: 10.1007/s11033-012-2097-7. PubMed PMID: 23065275.
554. Yildirim Y, Orhan EK, Iseri SA, Serdaroglu-Oflazer P, Kara B, Solakoglu S, et al. A frameshift mutation of ERLIN2 in recessive intellectual disability, motor dysfunction and multiple joint contractures. *Human molecular genetics*. 2011;20(10):1886-92. Epub 2011/02/19. doi: 10.1093/hmg/ddr070. PubMed PMID: 21330303.

555. Morava E, Kuhnisch J, Drijvers JM, Robben JH, Cremers C, van Setten P, et al. Autosomal recessive mental retardation, deafness, ankylosis, and mild hypophosphatemia associated with a novel ANKH mutation in a consanguineous family. *The Journal of clinical endocrinology and metabolism*. 2011;96(1):E189-98. Epub 2010/10/15. doi: 10.1210/jc.2010-1539. PubMed PMID: 20943778.
556. Saadi A, Borck G, Boddaert N, Chekkour MC, Imessaoudene B, Munnich A, et al. Compound heterozygous ASPM mutations associated with microcephaly and simplified cortical gyration in a consanguineous Algerian family. *European journal of medical genetics*. 2009;52(4):180-4. Epub 2009/04/01. doi: 10.1016/j.ejmg.2009.03.013. PubMed PMID: 19332161.
557. Bond J, Scott S, Hampshire DJ, Springell K, Corry P, Abramowicz MJ, et al. Protein-truncating mutations in ASPM cause variable reduction in brain size. *American journal of human genetics*. 2003;73(5):1170-7. Epub 2003/10/24. doi: 10.1086/379085. PubMed PMID: 14574646; PubMed Central PMCID: PMC1180496.
558. Bugiani M, Gyftodimou Y, Tsimpouka P, Lamantea E, Katzaki E, d'Adamo P, et al. Cohen syndrome resulting from a novel large intragenic COH1 deletion segregating in an isolated Greek island population. *American journal of medical genetics Part A*. 2008;146a(17):2221-6. Epub 2008/07/26. doi: 10.1002/ajmg.a.32239. PubMed PMID: 18655112.
559. Garshasbi M, Motazacker MM, Kahrizi K, Behjati F, Abedini SS, Nieh SE, et al. SNP array-based homozygosity mapping reveals MCPH1 deletion in family with autosomal recessive mental retardation and mild microcephaly. *Human genetics*. 2006;118(6):708-15. Epub 2005/11/29. doi: 10.1007/s00439-005-0104-y. PubMed PMID: 16311745.
560. Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, Pellerin S, et al. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *The New England journal of medicine*. 2009;360(6):599-605. Epub 2009/02/07. doi: 10.1056/NEJMoa0805392. PubMed PMID: 19196676; PubMed Central PMCID: PMC12925262.
561. Alazami AM, Al-Saif A, Al-Semari A, Bohlega S, Zlitni S, Alzahrani F, et al. Mutations in C2orf37, encoding a nucleolar protein, cause hypogonadism, alopecia, diabetes mellitus, mental retardation, and extrapyramidal syndrome. *American journal of human genetics*. 2008;83(6):684-91. Epub 2008/11/26. doi: 10.1016/j.ajhg.2008.10.018. PubMed PMID: 19026396; PubMed Central PMCID: PMC12668059.
562. Alazami AM, Schneider SA, Bonneau D, Pasquier L, Carecchio M, Kojovic M, et al. C2orf37 mutational spectrum in Woodhouse-Sakati syndrome patients. *Clinical genetics*. 2010;78(6):585-90. Epub 2010/05/29. doi: 10.1111/j.1399-0004.2010.01441.x. PubMed PMID: 20507343.
563. Habib R, Basit S, Khan S, Khan MN, Ahmad W. A novel splice site mutation in gene C2orf37 underlying Woodhouse-Sakati syndrome (WSS) in a consanguineous family of Pakistani origin. *Gene*. 2011;490(1-2):26-31. Epub 2011/10/04. doi: 10.1016/j.gene.2011.09.002. PubMed PMID: 21963443.
564. Ben-Omran T, Ali R, Almureikhi M, Alameer S, Al-Saffar M, Walsh CA, et al. Phenotypic heterogeneity in Woodhouse-Sakati syndrome: two new families with a

- mutation in the C2orf37 gene. *American journal of medical genetics Part A*. 2011;155a(11):2647-53. Epub 2011/10/04. doi: 10.1002/ajmg.a.34219. PubMed PMID: 21964978.
565. Guven A, Gunduz A, Bozoglu TM, Yalcinkaya C, Tolun A. Novel NDE1 homozygous mutation resulting in microhydranencephaly and not microlyssencephaly. *Neurogenetics*. 2012;13(3):189-94. Epub 2012/04/25. doi: 10.1007/s10048-012-0326-9. PubMed PMID: 22526350.
566. Ropers F, Derivery E, Hu H, Garshasbi M, Karbasiyan M, Herold M, et al. Identification of a novel candidate gene for non-syndromic autosomal recessive intellectual disability: the WASH complex member SWIP. *Human molecular genetics*. 2011;20(13):2585-90. Epub 2011/04/19. doi: 10.1093/hmg/ddr158. PubMed PMID: 21498477.
567. Cohn DH, Ehtesham N, Krakow D, Unger S, Shanske A, Reinker K, et al. Mental retardation and abnormal skeletal development (Dyggve-Melchior-Clausen dysplasia) due to mutations in a novel, evolutionarily conserved gene. *American journal of human genetics*. 2003;72(2):419-28. Epub 2002/12/20. doi: 10.1086/346176. PubMed PMID: 12491225; PubMed Central PMCID: PMC1420018.
568. Preiksaitiene E, Mannik K, Dirse V, Utkus A, Ciuladaite Z, Kasnauskiene J, et al. A novel de novo 1.8 Mb microdeletion of 17q21.33 associated with intellectual disability and dysmorphic features. *European journal of medical genetics*. 2012;55(11):656-9. Epub 2012/07/31. doi: 10.1016/j.ejmg.2012.07.008. PubMed PMID: 22842074.