

APPLICATION OF GENOMIC TECHNOLOGIES FOR MOLECULAR DIAGNOSIS OF GENETIC DISEASES



Karl Alexander Zhivojin Hudspith

St Peter's College
University of Oxford

*A thesis submitted in July 2015 for the degree of
Doctor of Philosophy*

Abstract

Methods for sequencing deoxyribonucleic acid (DNA) have improved rapidly in the past decade. Recent methods, termed “next-generation sequencing” (NGS), have made sequencing large quantities of DNA economically viable in molecular diagnostics of genetic diseases. This thesis describes some of the first investigations to use NGS for such a purpose.

Several different NGS platforms, each of which differs substantially in terms of sample preparation, chemistry, and sequencing methodology, were tested, in addition to several sequence capture and enrichment technologies that allow sequencing to be targeted, and these combinations were compared to Sanger sequencing. They were each found to have different strengths and weaknesses, which affect their accuracy, reliability, time taken to results, financial cost, and ease of use, but all showed high accuracy and dramatically increased throughput over Sanger sequencing.

NGS was used to identify pathogenic mutations in two groups of patients who had either inherited retinal dystrophies (IRD), or severe early onset epilepsies. NGS was able to identify pathogenic variants, demonstrating the ability of the technology to provide medically useful information for genetically heterogeneous conditions. NGS combined with a novel data analysis pipeline was able to make a secure molecular diagnosis in 25% of a cohort of IRD patients who previously did not have a genetic diagnosis.

The greatest challenge presented by NGS was found to be filtering the vast amounts of data produced to identify potential pathogenic variants. *In silico* pathogenicity prediction programs were used, but none were 100% accurate. Other methods were also employed to provide further evidence of pathogenicity. These included family based DNA testing for cosegregation of variants with phenotype, and transcript based analysis. In some patients, despite extensive genetic testing, a secure molecular diagnosis could not be made using DNA sequencing technologies, illustrating that challenges still remain in the field of genetic diagnostics.

Acknowledgements

I have had the good fortune to be mentored by five different supervisors over the course of my doctoral studies. I would like to start by thanking Jenny Taylor and Ioannis Ragoussis for initially admitting me to Oxford as a graduate student at the Wellcome Trust Centre for Human Genetics four-and-a-half-years ago. It might be a cliché phrase to use, but it really does feel like a lifetime ago.

Andrea Neméth only officially became my supervisor two-and-a-half-years in, but has been the one constant throughout my time here, and I am grateful for her support and belief in me when I made the decision to move to the Nuffield Department of Clinical Neurosciences, and turn what was originally going to be a one year masters into a four-and-a-half-year doctoral degree. My unplanned and slightly unorthodox course through graduate school caused a few extra problems on top of the many that scientific research naturally throws up. I would like to thank Andrea for not losing faith in me through some fairly turbulent days.

Stephanie Halford and Susan Downes both joined later as my supervisors, and have both been generous in sharing their time and expertise with me as well. I am very grateful for the time they have taken to provide with me feedback from experiments, talks I have given, and of course the long and arduous task of writing this thesis.

I have learned a great deal from all the people who I have worked with and been taught by, and another thanks goes to Morag Shanks and Alistair Pagnamenta at the WTCHG, both of whom I spent a lot of time collaborating with on a daily basis. There have been many others who I have worked with as well during my time here; Kulvinder Kaur, Leo Goodstadt, Stefano Lise, Richard Copley, and Hilary Martin, just to name a few. I'm grateful for all they have done to help me, including on many occasions taking my experiments off the thermocycler when they were finished so I could get to training on time.

Genetics is something that has fascinated me since I was very young, when I first started reading popular science books on the subject. It is a field of science that answers some of the most important questions about who we are, and what we might become.

My initial interest in science in general can most likely be traced back to my father, who would regularly take me on trips to the Science Museum and Natural History Museum when I was a child. I owe an awful lot to both my parents for the support they have given me, not only during my time here at Oxford, but in all the years before then as well.

Finally, while I have learned a lot in the lab, I have also received a well-rounded education at the university through my time in the Oxford University Boat Club. My friends are almost exclusively rowers, and many of my team-mates have been scientists as well. The ability to talk to them for advice, support, and general lifting of morale when experiments have failed to work, or other problems have occurred, has helped to get me through the past four-and-a-half-years. I would like to thank in particular Malcolm Howard, Alex Woods, Tom Watson, and Alex Davidson, all of whom are fellow molecular biologists.

I am also grateful for the mentoring I have received from my coach, Sean Bowden. The curriculum of rowing is more limited than that of science, but it teaches the same necessary requirements for success; teamwork, leadership, creativity, courage, integrity, resolve, trust, communication, resilience, focus, determination, and above all, hard work.

Thank you to everyone who has helped me get to the finish line.

Statement of originality

I declare that this thesis is my own work, and that data reported are from experiments performed by me unless stated below:

- Next-generation sequencing data in Chapter 3. I performed the initial fragmentation and library preparation procedures for Illumina sequencing as part of the experiments described in Chapter 3. The technology was introduced in Chapter 1 section 1.3.2. Full details of the methods I used can be found in Chapter 3 section 3.2.1. All remaining steps of the process were performed by the High Throughput Sequencing Group at the Wellcome Trust Centre for Human Genetics (University of Oxford). Sanger sequencing was performed by the Oxfordshire NHS Regional Genetics Laboratory. All the remaining laboratory experiments from which data was derived in this chapter were performed by Drs Morag Shanks and Alix Kwasniewska. Data processing from raw data to generate a list of variants was performed by Dr Richard Copley. All analysis and filtering on these data from this point was performed by me.
- I performed all stages of the Roche 454 GS FLX sequencing method for the experiments described in Chapter 4. This technique was introduced in Chapter 1 section 1.3.1. A summary of the method I used is given in Chapter 4 section 4.2.1. The full methods can be found in the appendix section 9.1. Lab work was a collaboration between myself and Dr Morag Shanks who also performed all of the sequence capture and enrichment steps. Data processing from raw data to generate a list of variants was performed by Drs Richard Copley and Stefano Lise. All analysis and filtering on these data from this point was performed by me.
- All interactions with patients, including diagnosis of clinical phenotypes and genetic counselling were performed by Drs Andrea Nemeth and Susan Downes.

All remaining work in the thesis that was performed by others, or figures used from other sources, are stated within the thesis at the relevant points. Work that is not otherwise attributed to another source was performed by me.

Table of contents

Abstract	2
Acknowledgements	3
Statement of originality	5
Table of contents	6
List of figures	11
List of tables	14
Abbreviations	15
CHAPTER 1	18
Introduction	18
1.1 Mutations and genetic variation	19
1.1.1 Effects of mutations	20
1.2 Technologies for sequencing DNA	24
1.2.1 Sanger sequencing	24
1.3 Next-generation Sequencing (NGS)	28
1.3.1 Emulsion based PCR amplification and pyrosequencing	28
1.3.2 Sequencing by synthesis using reversible terminator chemistry	32
1.3.3 Whole genome sequencing using unchained base reads on self-assembling DNA nanoballs	35
1.4 Sequencing costs and sequence capture	37
1.5 DNA sequence capture and enrichment technologies	39
1.5.1 NimbleGen sequence capture array	39
1.5.2 Agilent SureSelect solution capture	40
1.5.3 RainDance microdroplet-based PCR	41
1.6 Bioinformatic analysis of sequencing data	42
1.6.1 Evaluating alignment	43
1.6.2 Coverage of bases	43
1.6.3 Variant calling and analysis	45
1.6.4 Creating a bioinformatic pipeline	48
1.7 Clinical utility of genetic testing	51
1.7.1 Overview of genetic technologies used in diagnostics	51
1.8 Genetic disorders studied in this project	54
1.8.1 Inherited retinal dystrophies (IRD)	54
1.8.2 Genetic testing for retinal disorders	65

1.9 Clinical utility of next-generation sequencing with sequence capture	67
1.9.1 Aims	69
CHAPTER 2	70
Materials and Methods	70
2.1 Materials	71
2.1.1 Enzymes	71
2.1.2 Solutions	71
2.1.3 Other materials	71
2.1.4 Equipment and apparatus	72
2.1.4.1 Thermocycler	72
2.1.4.2 Electrophoresis	72
2.1.4.3 Bioanalyzer	72
2.1.4.4 Sanger sequencing	72
2.1.4.5 Next-Generation Sequencing	72
2.1.4.6 Sequence capture technologies	72
2.1.4.7 Spectrophotometry equipment	73
2.1.1.8 SequalPrep Normalization Plates	73
2.1.5 Software	73
2.2 Methods	75
2.2.1 Patient cohorts and ophthalmological examination	75
2.2.2 DNA extraction and dilution	75
2.2.3 Nucleic acid quantitation	76
2.2.3.1 Spectrophotometry using the Nanodrop apparatus	77
2.2.3.2 Spectrophotometry using the Qubit fluorometric quantitation method	77
2.2.4 Agarose gel electrophoresis	78
2.2.5 Agilent 2100 Bioanalyzer	78
2.2.6 Polymerase Chain Reaction (PCR)	79
2.2.6.1 Primer design	80
2.2.6.2 Standard PCR	82
2.2.6.3 Long Range PCR with Illumina HiSeq 2500 sequencing	83
2.2.7 DNA purification	86
2.2.7.1 QIAquick PCR Purification Kit	86
2.2.7.2 Solid Phase Reversible Immobilisation (SPRI) beads	86
2.2.8 Sequencing	87
2.2.8.1 Sanger sequencing	87
2.2.8.2 Sequence capture and enrichment	90

2.2.8.3 Next-generation sequencing	94
CHAPTER 3	95
Comparison of Next-generation Sequencing with Sanger sequencing	95
3.1 Introduction	96
3.2 Materials and Methods	98
3.2.1 Illumina sequencing	99
3.2.2 Data analysis	104
3.3 Results	105
3.3.1 Overall variant detection rates in Capture/NGS compared to Sanger sequencing in 24 patients	105
3.3.2 Alignment of sequence capture technologies	109
3.3.3 Other factors affecting selection of a sequence capture technology	110
3.3.4 Coverage of NGS platforms	112
3.3.5 Comparison of “ease of experimental use” of each NGS platform	115
3.4 Discussion	118
CHAPTER 4	121
Investigating the clinical utility of NGS in patients with Inherited Retinal Dystrophies	121
4.1 Introduction	122
4.2 Materials and Methods	125
4.2.1 Summary of NimbleGen sequence capture and 454 sequencing process	126
4.2.2 Experimental quality control data	127
4.2.3 Exome sequencing	143
4.3 Results	143
4.3.2 Development of bioinformatic data analysis pipeline	143
4.3.3 Roche 454 sequencing results	148
4.3.4 Pathogenic variants identified	149
4.3.5 Exome sequencing analysis of patient 525A6	155
4.4 Discussion	160
4.4.1 Conclusions	163
CHAPTER 5	164
Testing variants identified by whole genome sequencing in patients with early onset epilepsy	164
5.1 Introduction	165
5.1.1 Early onset epilepsy	166
5.1.2 Preliminary analysis of identified variants in patients OTH 12 and OTH 18	169
5.2 Materials and Methods	173

5.2.1 mRNA extraction from blood	173
5.2.2 Reverse transcription and amplification of cDNA	174
5.3 Results	175
5.3.1 Validation and cosegregation of variants	175
5.3.2 Determination of splicing effect in mRNA	183
5.4 Discussion	188
5.4.1 Mutations in <i>PIGQ</i> can cause Ohtahara syndrome	188
5.4.2 A variant in <i>CBL</i> causes a range of phenotypes	190
5.4.3 Conclusion	191
CHAPTER 6	193
Searching for pathogenic variants in macular dystrophies	193
6.1 Introduction	194
6.1.1 Detection of mutations in Stargardt disease with different technologies	203
6.1.2 Screening of the <i>ABCA4</i> gene in STGD patients in the UK	204
6.1.3 Aims of the work presented in this chapter	208
6.1.4 Hypotheses to account for why a second pathogenic variant in <i>ABCA4</i> cannot be identified in some patients with STGD	209
6.1.5 Investigations that were performed in the work described in this chapter	214
6.2 Methods and materials	219
6.2.1 Targeted next-generation sequencing	220
6.2.2 Whole genome sequencing using ligation based sequencing and DNA nanoballs	221
6.2.3 Long Range PCR generated amplifications of the <i>ABCA4</i> gene, sequenced on the Illumina HiSeq 2500	222
6.2.4 Reverse transcription and amplification of <i>ABCA4</i> cDNA	224
6.3 Results	227
6.3.1 Targeted sequencing of other retinal genes	227
6.3.2 Re-analysis of targeted sequencing data to search for variants in cryptic exon in <i>ABCA4</i> identified by Braun and colleagues	231
6.3.3 Long-Range PCR amplification of the <i>ABCA4</i> genomic region	234
6.3.5 Whole genome sequencing of ten patients with Stargardt disease and a single potentially pathogenic variant identified in the <i>ABCA4</i> gene	241
6.3.6 <i>ABCA4</i> cDNA analysis in patients using mRNA from lymphoblastoid cell lines	245
6.4 Discussion	247
6.4.1 Further work and conclusion	252

CHAPTER 7	255
General discussion	255
7.1. The progression of DNA sequencing technologies	259
7.2 Data analysis	263
7.2.1 Further evidence of pathogenicity	266
7.2.2 Reference Genomes	267
7.3 Remaining Challenges	270
7.3.1 Third generation sequencing technologies	271
7.4 The future of genomics	272
7.5 Conclusion	275
8. References	276
9. Appendix	293
9.1 454 GS FLX sequencing methods	294
9.1.1 454 DNA library preparation	294
9.1.2 Array capture	299
9.1.3 Emulsion PCR	302
9.1.4 Sequencing	308
9.2 Cohort of patients with macular dystrophies	314
10. Publications	317

List of figures

Figure 1.1 Different types of mutations caused by single nucleotide variants (SNVs).	23
Figure 1.2 Example of a small deletion and the subsequent frameshift effect it causes.	23
Figure 1.3 Structural formula of a deoxynucleotide and a dideoxynucleotide.	26
Figure 1.4 The original dideoxy chain-termination reaction method.	26
Figure 1.5 Basic principle of shotgun sequencing.	27
Figure 1.6 454 DNA Library preparation process.	29
Figure 1.7 454 emulsion PCR process.	30
Figure 1.8 454 pyrosequencing.	31
Figure 1.9 Illumina sequencing part 1.	33
Figure 1.10 Illumina sequencing part 2.	34
Figure 1.11 Complete Genomics sequencing technology.	36
Figure 1.12 Cost of genome sequencing over time.	38
Figure 1.13 Schematic diagram of the NimbleGen sequence capture array workflow.	39
Figure 1.14 Schematic diagram of the SureSelect workflow.	40
Figure 1.15 RainDance microfluidics technology.	41
Figure 1.16 Alignment of NGS reads.	42
Figure 1.17 Simplified example of how shotgun sequencing data is compiled <i>in silico</i> .	44
Figure 1.18 Bioinformatic pipeline.	50
Figure 1.19 The mammalian eye.	59
Figure 1.20 A molecular graphics ribbon diagram of the structure of the rhodopsin protein.	60
Figure 1.21 Major components of the rod phototransduction and retinoid cycles.	61
Figure 1.22 Fundus image of a normal retina.	62
Figure 1.23 Retinal images of retinitis pigmentosa.	63
Figure 1.24 Images of Stargardt disease.	63
Figure 1.25 Imaging of patient with cone-rod dystrophy.	64
Figure 1.26 Autofluorescence image of the right eye of a patient with cone-rod dystrophy.	64
Figure 1.27 The number of mapped and identified retinal disease genes between 1980 and 2014.	66
Figure 3.1 GS Reference Mapper software showing a miscalled variant in a homopolymer tract.	107
Figure 4.1 Flowchart of 454 sequencing methodology in conjunction with NimbleGen sequence capture array.	126
Figure 4.2 DNA nebulisation quality control data.	129
Figure 4.3 Quality control data for DNA library preparations.	130
Figure 4.4 Quality control data for pre-capture LM-PCR amplified DNA libraries.	131

<u>Figure 4.5</u> Image of wells producing data that passed quality control data filtering.	134
<u>Figure 4.6</u> Percentages of library reads that passed quality control filtering.	135
<u>Figure 4.7</u> Homopolymer signals in a sequencing run.	137
<u>Figure 4.8</u> Control DNA metrics.	137
<u>Figure 4.9</u> Read lengths in an average sequencing run.	139
<u>Figure 4.10</u> Raw data analysis using the GS Reference Mapper software.	141
<u>Figure 4.11</u> Example of a flowgram on the GS Reference Mapper software.	142
<u>Figure 4.12</u> Initial data analysis pipeline.	145
<u>Figure 4.13</u> Improved data analysis pipeline.	147
<u>Figure 4.14</u> Screenshot of Ingenuity Variant Analysis results for congenital myasthenic syndrome in patient 525A6.	158
<u>Figure 4.15</u> Screenshot of Ingenuity Variant Analysis results for retinitis pigmentosa in patient 525A6.	159
<u>Figure 5.1</u> Electroencephalogram (EEG) trace showing a burst suppression pattern consistent with Ohtahara syndrome.	167
<u>Figure 5.2</u> Pedigrees of the early onset epilepsy families.	172
<u>Figure 5.3</u> Confirmation of a heterozygous c.4367C>T substitution in the <i>C5orf42</i> gene in patient OTH 12.	177
<u>Figure 5.4</u> Identifying the phase of a variant using IGV.	178
<u>Figure 5.5</u> Segregation of the <i>SCN4A</i> variant in the family of OTH 12.	179
<u>Figure 5.6</u> Segregation of the exon 3 c.690-2A>G substitution in the <i>PIGQ</i> gene of the family of patient OTH 12.	180
<u>Figure 5.7</u> NGS data showing two <i>de novo</i> SNVs in patient OTH 18.	181
<u>Figure 5.8</u> Two SNVs in the <i>CBL</i> gene of patient OTH 18.	182
<u>Figure 5.9</u> Bioanalyzer data showing amplicons of cDNA from samples OTH 18, 20, 13, and 14.	184
<u>Figure 5.10</u> Exon skipping in the <i>PIGQ</i> gene in the parents of patient OTH 12.	185
<u>Figure 5.11</u> Exon skipping in the <i>CBL</i> gene in patient OTH 18.	187
<u>Figure 5.12</u> GPI biosynthesis pathway.	189
<u>Figure 6.1</u> Images of Stargardt disease in four patients.	196
<u>Figure 6.2</u> Various images of a bullseye maculopathy.	197
<u>Figure 6.3</u> Colour photograph showing chorioretinal atrophy.	197
<u>Figure 6.4</u> Composite image of a patient showing Fundus Flavimaculatus.	198
<u>Figure 6.5</u> Structure of the <i>ABCA4</i> gene.	200
<u>Figure 6.6</u> Structure of the <i>ABCA4</i> transporter protein.	201
<u>Figure 6.7</u> Correlation between severity of <i>ABCA4</i> mutations and phenotypes.	202
<u>Figure 6.8</u> Amplicons produced by the LR-PCR amplifications.	223

<u>Figure 6.9</u> Retinal imaging and pedigree of the family of patient 245F6.	229
<u>Figure 6.10</u> Cosegregation of the c.121C>T (p.Arg41Trp) <i>CRX</i> variant in the family of patient 245F6.	230
<u>Figure 6.11</u> Screenshot of the UCSC Genome Browser displaying the genomic region between the <i>GCLM</i> and <i>ARHGAP29</i> genes.	233
<u>Figure 6.12</u> Example gel electrophoresis image showing 18 Long Range PCR amplicons covering the <i>ABCA4</i> gene.	236
<u>Figure 6.13</u> Example screenshot from UCSC genome browser confirming an amplification maps to the correct place on the <i>ABCA4</i> gene.	237
<u>Figure 6.14</u> The potential effect of a rearrangement on paired end reads produced by NGS.	240
<u>Figure 6.15</u> Screenshots of the UCSC genome browser displaying data from the retinal transcriptome paper by Farkas <i>et al.</i> (2013).	244
<u>Figure 6.16</u> Successful nested PCR amplifications of <i>ABCA4</i> cDNA.	246
<u>Figure 7.1</u> The cost of genome sequencing over the previous 14 years.	260
<u>Figure 7.2</u> Schematic for next-generation sequencing and variant detection as reported by Watson and colleagues (2014).	265
<u>Figure 7.3</u> Creating a more accurate reference sequence using hydatidiform moles.	269

List of tables

<u>Table 2.1</u> List of software used.	74
<u>Table 2.2</u> Validation primers.	81
<u>Table 2.3</u> <i>ABCA4</i> Long Range-PCR primers.	85
<u>Table 2.4</u> Genes used in targeted capture.	91
<u>Table 3.1</u> Detection rate of <i>ABCA4</i> variants.	106
<u>Table 3.2</u> Alignment of captured DNA reads to target design.	111
<u>Table 3.3</u> Average coverage summary.	114
<u>Table 3.4</u> Technical comparison of ease of experimental use of each NGS platform.	117
<u>Table 4.1</u> Summarised sequencing results for first batch of 28 patients analysed using initial pipeline.	148
<u>Table 4.2</u> Summarised sequencing results for next 22 patients analysed using second pipeline.	148
<u>Table 4.3</u> Pathogenic mutations identified through the analysis pipeline.	154
<u>Table 6.1</u> Patients diagnosed with a non-Stargardt disease macular dystrophy phenotype in whom only a single potentially pathogenic variant in the <i>ABCA4</i> gene has been identified.	206
<u>Table 6.2</u> Patients diagnosed with a Stargardt disease compatible phenotype in whom only a single potentially pathogenic variant in the <i>ABCA4</i> gene have been identified.	207
<u>Table 6.3</u> Experiments performed on the patients with macular dystrophy phenotypes.	218
<u>Table 6.4</u> <i>ABCA4</i> cDNA primers.	226
<u>Table 6.5</u> The seven pathogenic intronic or synonymous variants identified by Braun and colleagues (Braun <i>et al.</i> 2013).	232
<u>Table 6.6</u> Table of preliminary sequencing metrics data for 14 patients and 6 controls sequenced from LR-PCR amplicons.	239
<u>Table 6.7</u> Patients in whom a potential molecular diagnosis has been identified from the investigations in this chapter.	249
<u>Table 9.1</u> Patients diagnosed with a Stargardt disease compatible phenotype in whom at least two potentially pathogenic variants in <i>ABCA4</i> have been identified.	314
<u>Table 9.2</u> Patients with macular dystrophy phenotypes other than STGD or FF.	316

Abbreviations

μ	Micro (10 ⁻⁶)	ARVC	Arrhythmogenic right ventricular cardiomyopathy
A	Adenine		
aa	Amino acid	ATP	Adenosine triphosphate
ABCA4	ATP-binding cassette transporter alpha-4 subunit	BBS	Bardet-Biedl syndromes
ad	Autosomal dominant (usually used as a prefix e.g. adRP, adCRD)	BEM	Bull's eye maculopathy
ADP	Adenosine diphosphate	BGI	Beijing Genomics Institute
AF	Autofluorescence	bp	Base pair(s)
AFA	Adaptive focussed acoustics	C	Cytosine
AMD	Age-related macular degeneration	cDNA	Complementary DNA
APEX	Arrayed primer extension	CFCS	cardiofaciocutaneous syndrome
ar	Autosomal recessive (usually used as a prefix e.g. arRP, arCRD)	cGMP	Cyclic guanosine monophosphate
Array CGH	Array comparative genomic hybridisation	cM	Centimorgan
		CRD	Cone-rod dystrophy
		CRX	Cone-rod homeobox
		DCM	Familial dilated cardiomyopathy

DNA	Deoxyribonucleic acid	kbp	Kilo base pairs
dNTP	Deoxyribonucleotide	LCA	Leber congenital amaurosis
ds	Double stranded	LINE	Long interspersed nuclear elements
ELIDA	Enzymatic luminometric inorganic pyrophosphate detection assay	LM-PCR	Ligation mediated-PCR
emPCR	Emulsion PCR	m	Metre or milli (10^{-3})
ER	Endoplasmic reticulum	MAPP	Multivariate Analysis of Protein Polymorphism
EVS	Exome variant server	masterVar	Master variant file
Exo	Exonuclease	MD	Macular degeneration
FF	Fundus flavimaculatus	min	Minute(s)
g	Gram(s)	MLPA	Multiplex ligation probe amplification
G	Guanine	mRNA	Messenger RNA
gDNA	Genomic DNA	MID	Molecular identifier
HCM	Hypertrophic cardiomyopathy	miRNA	Micro RNA
HSP	Hereditary spastic paraplegia	n	Nano (10^{-9})
IGV	Integrative genomics viewer	NCBI	National Centre for Biotechnology Information
IRD	Inherited retinal dystrophies	NGS	Next-generation sequencing
k	Kilo (10^3)		

OD	Optical density	SSCP	Single strand conformation polymorphism
ORF	Open reading frame		
OTH	Ohtahara syndrome	SPRI	Solid phase reversible immobilisation
p	Pico (10^{-12})	ss	Single stranded
PCR	Polymerase chain reaction	STGD	Stargardt disease
PNK	Polynucleotide 5'-hydroxyl-kinase	T	Thymine
PPI	Pyrophosphate ($P_2O_7^{4-}$)	T _a	Annealing temperature
qPCR	Quantitative PCR	TAE	Tris-Acetate-EDTA
RefSeq	Reference sequence	TE	Tris-EDTA
RNA	Ribonucleic acid	T _m	Melting temperature
RP	Retinitis pigmentosa	TRE	Trinucleotide repeat expansion
RPE	Retinal pigment epithelium	UCSC	University of California Santa Cruz
rpm	Revolutions per minute	UTR	Untranslated region
RT PCR	Reverse transcription PCR	VCF	Variant called file
SAP	Shrimp alkaline phosphatase	W	Watt(s)
SFF	Standard flowgram format	wt	Wild type
SNP	Single nucleotide polymorphism		
SNV	Single nucleotide variant		

CHAPTER 1

Introduction

1.1 Mutations and genetic variation

The DNA in the genomes of any two individuals will differ due to natural genetic variation within a species, which accounts for factors such as eye colour, hair colour, and height. DNA replication includes a proof-reading ability, so the majority of the DNA on a chromosome remains unchanged between cell divisions.

Mutations are changes to DNA sequence. They often occur during cell divisions, and if present in the germline they can be inherited by the next generation. Mutations can be benign, beneficial, or pathogenic.

Care must be taken when describing mutations and sequence variants to ensure that a description of a sequence change does not incorrectly imply pathogenicity or even whether such a determination has been made. Even describing variants in DNA can present difficulties, as every variant is only a variant relative to another genome. There is no one single genome sequence that can act as a definitive version of the human genome, even though the reference sequence (RefSeq) provided by the Genome Reference Consortium (<http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/>) is often treated by molecular biologists as such.

An example of this difficulty is provided in the definition of polymorphism by Strachan and Read:

“Polymorphism – Strictly, the existence of two or more variants (alleles, phenotypes, sequence variants, chromosomal structure variants) at significant frequencies in the population. Looser usages among molecular geneticists include (1) any sequence variant present at a frequency >1% in a population (2) any non-pathogenic sequence variant, regardless of frequency.” (Strachan and Read 2011).

This then leads to further confusion with the commonly used term single nucleotide polymorphism (SNP):

“A position in the genome where two or occasionally three alternative nucleotides are common in the population. May be pathogenic or neutral. The dbSNP database lists human SNPs, but includes some rare pathogenic variants and some variants that involve two or more contiguous nucleotides.” (Strachan and Read 2011).

To avoid confusion, for the purposes of this thesis the term “polymorphism” will be used to describe variants or other mutations that are considered to be non-pathogenic or benign from a disease-causing standpoint. These alleles tend to exist at high frequency in the general population.

The term single nucleotide variant (SNV) will be used to refer to any variation affecting just a single nucleotide found in an individual by comparison with the human genome reference sequence. Use of this term does not imply whether the variant is pathogenic or not.

The term single nucleotide polymorphism (SNP) will be used to refer to SNVs which are presumed to be polymorphisms (as defined above).

1.1.1 Effects of mutations

Mutations that occur in genomic DNA can cause a disease phenotype through a variety of mechanisms. One of the primary purposes of DNA is to code for proteins, so DNA sequence variants that occur in protein-coding regions will often have an effect on function.

SNVs that are found in codons can be classified according to the effect they have on translation, as shown in **Figure 1.1**. Due to the degenerate nature of amino acid translation some mutations are synonymous and so do not change the amino acid that the codon transcribes, whereas non-synonymous variants will change just the single amino acid translated by the codon they are located in. A non-synonymous variant that changes just one amino acid can cause pathogenicity if the amino acid is part of a functionally important domain of a protein, such as the active site of an enzyme, or a cell surface receptor.

Nonsense and readthrough mutations have a greater effect as they respectively halt or continue the translation of all codons downstream of them. Frameshifts caused by insertions or deletions that are not evenly divisible by three can also have a severe effect on protein function as shown in **Figure 1.2**. For example, shortly after the work in this thesis began, a 15 bp deletion causing a frameshift that created a premature stop codon in the scaffold protein containing Ankyrin repeats (*USH1G*) gene was reported to cause Usher syndrome in a family (Bashir *et al.* 2010).

The mRNA transcripts produced from genes containing nonsense, readthrough, or frameshift mutations will often be destroyed by nonsense-mediated decay, an error detection process the cell uses to prevent the creation of faulty proteins (Chang *et al.* 2007).

Mutations to the essential splice sites of introns can also cause a disease phenotype by preventing splicing from occurring correctly, which can result in an exon being skipped. Mutations in splice sites have been reported to cause a range of retinal disorders. For example a homozygous single nucleotide variant (c.187+1G>T) in the donor splice site of intron 3 of the gamma subunit of the rod photoreceptor cyclic GMP-phosphodiesterase (*PDE6G*) gene was recently reported to cause retinitis pigmentosa in a family (Dvir *et al.* 2010). The variant causes the skipping of exon three when assayed in a mutant construct *in vitro* (Dvir *et al.* 2010).

Mutations can be the result of intrinsic errors during replication and repair processes and can also be induced by extrinsic mutagens such as radiation or carcinogenic chemicals. They can occur during both meiosis (in the germline) and mitosis. If a mutation occurs during meiosis and is passed on to the embryo it is said to be *de novo*, in contrast to mutations that are inherited from the parents.

The human genome is diploid; other than the sex chromosomes, there are two copies of every chromosome, and therefore two copies of each gene, one maternal and one paternal. In some genetic diseases a mutation of only one copy of the gene causes the disease phenotype; such conditions are referred to as having an autosomal dominant inheritance pattern because of the two alleles at a single locus only one is mutated.

Conditions in which both alleles are mutated to cause the disease are known as autosomal recessive. In autosomal recessive conditions the parents of affected offspring are often carriers who are healthy themselves, but carry one defective copy of the gene each.

The genetics of X-linked disorders (one of the two sex chromosomes) are more complicated, as males have only one copy of the X chromosome, whereas females have two.

Mutation type	synonymous			nonsynonymous			nonsense			readthrough		
Mutant amino acid	Ser			Asp			Stop			Gln		
Mutant mRNA sequence	UCG			GAC			UAA			CAA		
Original gDNA sequence	TCACAAAAGCCTCGTATTCAATTGGCTAAGGCATAT.											
Original amino acid	Ser	Gln	Lys	Ala	Ser	Tyr	Ser	Ile	Gly	Stop		

Figure 1.1 Different types of mutations caused by single nucleotide variants (SNVs).

Mutant amino acid	Ser	Lys	Ser	Leu	Val	Phe	Asn	Trp	Leu	Arg	His
Mutant mRNA sequence	UCAAAAAGCCUCGUAUUCAAUUGGCUAAGGCAUA..										
Original gDNA sequence	TCA CA AAAAGCCTCGTATTCAATTGGCTAAGGCATA..										
Original amino acid	Ser	Gln	Lys	Ala	Ser	Tyr	Ser	Ile	Gly	Stop	

Figure 1.2 Example of a small deletion and the subsequent frameshift effect it causes.

1.2 Technologies for sequencing DNA

Early methods for sequencing DNA were extremely tedious and it took several weeks to sequence just a few base pairs. The great breakthrough came from Frederic Sanger in Cambridge, whose method of di-deoxy chain termination allowed for the rapid sequencing of hundreds of base pairs at a time (Sanger and Coulson 1975, Sanger *et al.* 1977). Sanger sequencing, as it subsequently became known, quickly became the method of choice for gene sequencing projects and later for clinical diagnostics (Sanger *et al.* 1977).

1.2.1 Sanger sequencing

This method of sequencing utilises the chemical structure of dideoxynucleotides which lack the 3' OH group to which another nucleotide can form a bond (**Figure 1.3**).

The original Sanger method used four reaction mixtures containing one of the four modified dideoxynucleotides in addition to all other reagents (including deoxynucleotides) necessary to synthesise a complementary DNA strand. The presence of dideoxynucleotides in the reaction mix terminates elongation of the complementary DNA strand. Since this can occur at various positions in the molecule, the terminated fragments are of different lengths. The strands are separated by size on a polyacrylamide gel, with each reaction mixture run out on a separate lane next to each other. Visualisation of the DNA fragments is enabled by radioactively labelling the dideoxynucleotides, allowing an image of the gel to be made using X-ray film (**Figure 1.4**).

Over the years since the method was first described, numerous improvements have been made, the most recent being the use of fluorescent markers instead of radioactive labelling of the dideoxynucleotides, and the use of capillary sequencers to separate the fragments rather than gel electrophoresis.

Sanger sequencing is still commonly used in molecular biology, and is considered to be the gold standard for accuracy with reported error rates consistently less than 1% (Ewing *et al.* 1998). The main disadvantage of Sanger sequencing is that only 500-700 base pairs of sequence are produced in a single reaction. As a result, the human genome project cost \$2.7 billion to produce

a draft sequence of the human genome, and took twelve and a half years to complete. Routinely sequencing genomes through this method would be prohibitively time consuming and expensive for other uses such as clinical diagnostics. Therefore, different approaches have been developed, known as “next-generation sequencing” (NGS) or “high-throughput sequencing”.

Most high-throughput approaches to sequencing large amounts of DNA rely on a method of DNA sequencing known as shotgun sequencing in which the DNA is fragmented into smaller molecules that can be sequenced individually and in parallel.

The individual fragments are known as a “library” and may be individually amplified (by cloning or PCR) as an additional step. The library therefore represents the original DNA source which could be genomic DNA or an enriched subset of DNA molecules.

Each of these fragments is sequenced in a separate reaction, and the sequence reads are then re-compiled into a continuous sequence on a computer by aligning them where they overlap (**Figure 1.5**).

All next-generation sequencing methods so far developed rely on shotgun sequencing. They increase the throughput compared to Sanger sequencing by performing millions of simultaneous sequencing reactions, sometimes known as “massively parallel sequencing”. Although they share this basic principle in common, NGS technologies differ greatly in other aspects of their methodologies and chemistries.

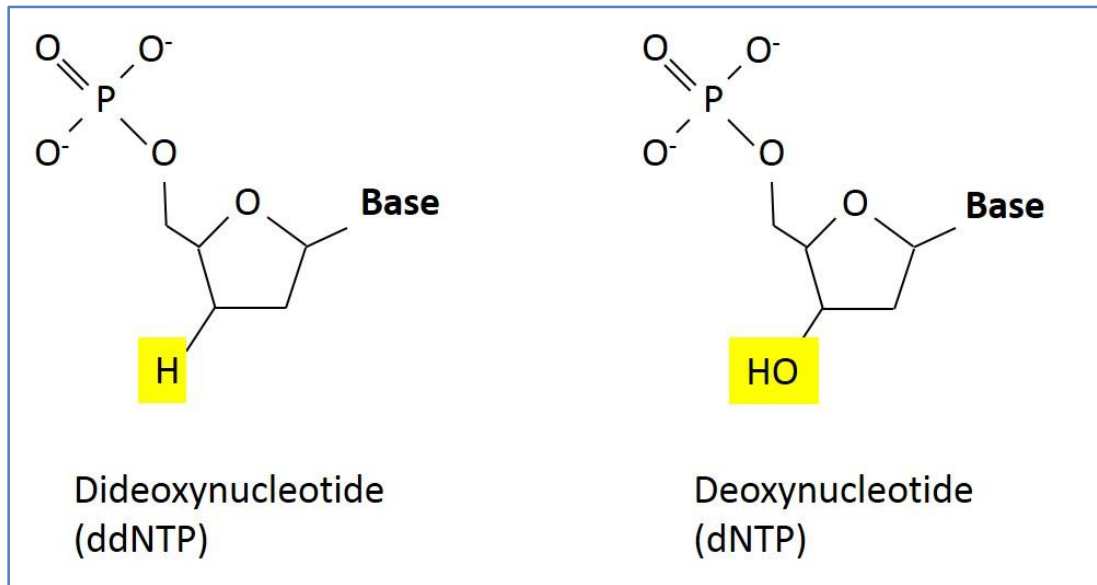


Figure 1.3 Structural formula of a deoxynucleotide and a dideoxynucleotide. Each nucleotide is made up of a pentose ring, a phosphate group, and a base. In deoxynucleotides there is an OH group on the 3' carbon of the pentose ring. In dideoxynucleotides there is an H instead, which prevents binding of the phosphate group of another nucleotide and the formation of a polymer.

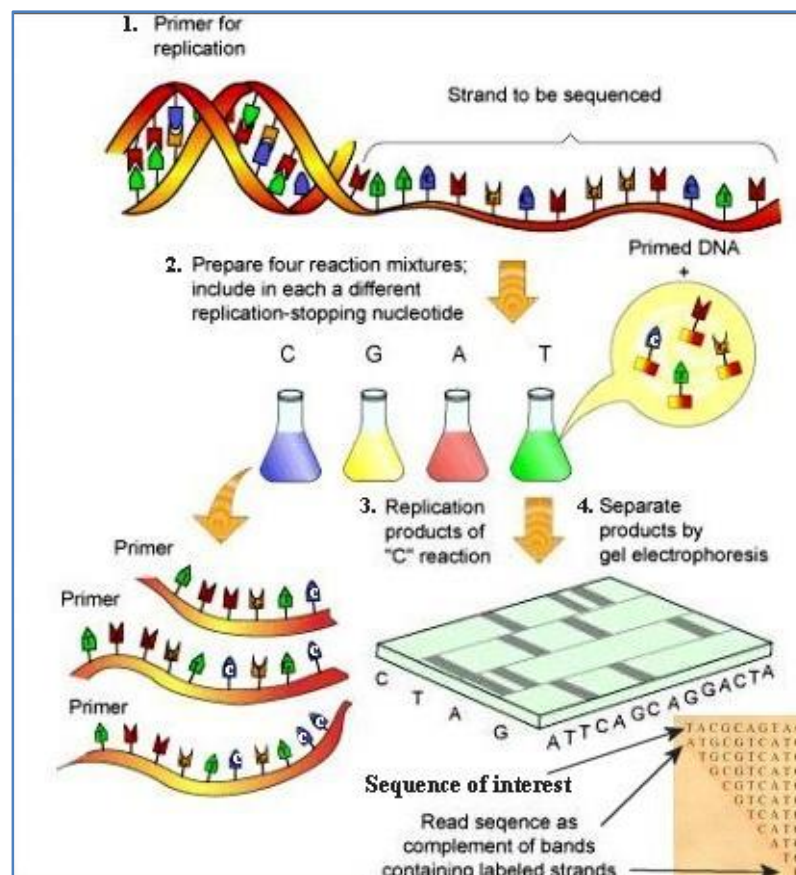


Figure 1.4 The original dideoxy chain-termination reaction method. Four individual reactions are performed, each with a different chain terminating dideoxynucleotide included, and the products generated from each reaction are separated on an agarose gel, allowing the identity of the nucleotide at each position of the sequence to be determined. Figure taken from <http://universe-review.ca/I11-50-DNAsequencing1.jpg>

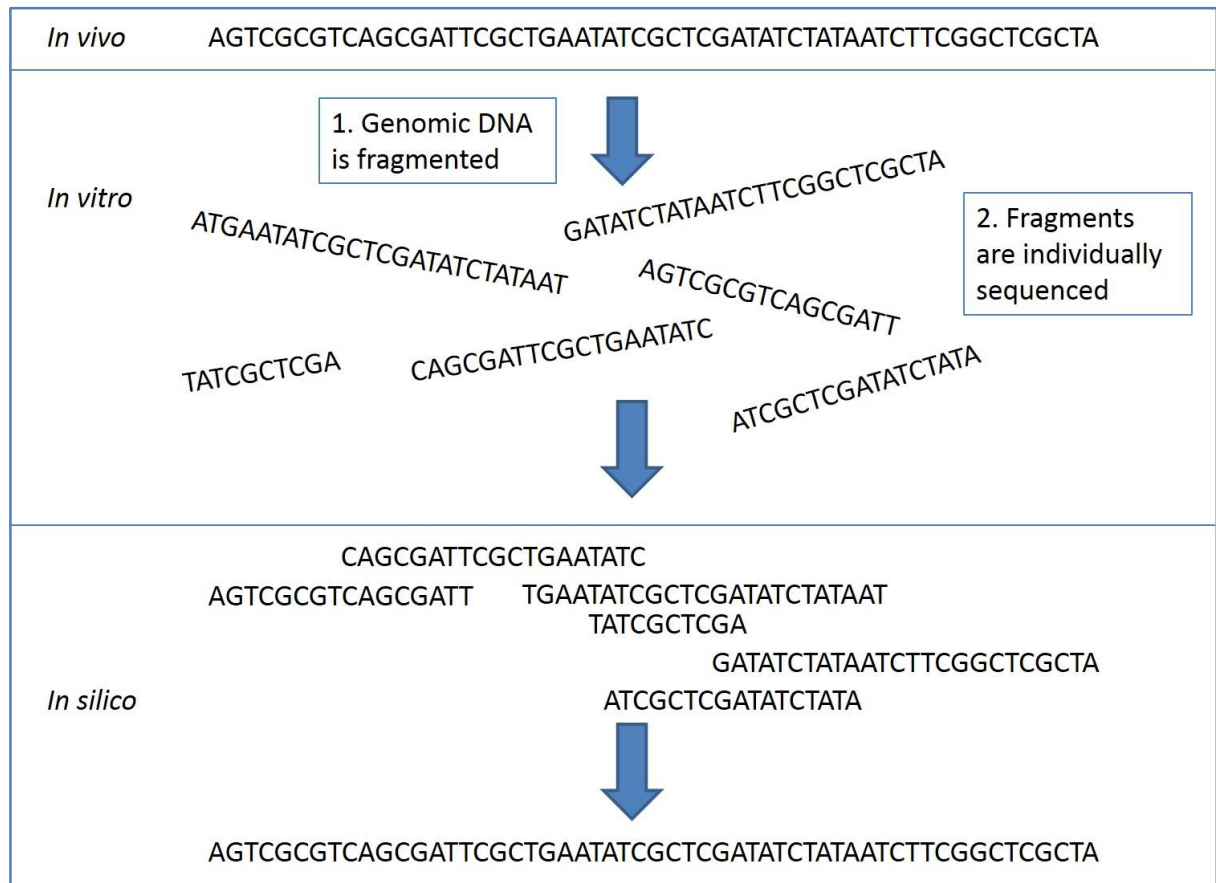


Figure 1.5 Basic principle of shotgun sequencing. DNA is first fragmented into small fragments which are then individually sequenced. Overlapping reads are overlaid to develop a continuous sequence using the human genome reference sequence (RefSeq) as a guide.

1.3 Next-generation Sequencing (NGS)

Several different Next-generation Sequencing (NGS) technologies have been developed by pioneering scientists in the field. The principles of the three technologies used in this project are discussed below.

1.3.1 Emulsion based PCR amplification and pyrosequencing

The first NGS platform was developed by Jonathon Rothberg in 1999. Rothberg utilised a number of different existing technologies to develop a method of rapidly amplifying and then sequencing many millions of DNA fragments in parallel (Margulies *et al.* 2005).

Firstly, DNA is prepared into shotgun DNA libraries by fragmentation using nitrogen gas, followed by the addition of adaptors to each end of the fragments. The adaptors are used to attach the many different fragments of the DNA library to individual microscopic polystyrene beads. These beads are emulsified in a water-in-oil emulsion, creating independent microreactors in which many millions of different DNA fragments can be amplified in parallel in an emulsion PCR (emPCR, Williams *et al.* 2006).

After emPCR the emulsion is broken and the clonally amplified fragments covering the beads are denatured. The beads are packaged into microfabricated wells on a device known as a picotiter plate (PTP) (Leamon *et al.* 2003) inside the sequencing machine, where a reaction based on the enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA) is performed (Nyren and Lundin 1985, Nyren and Edwin 1994). This involves the sequential cycling of nucleotides onto the surface of the PTP, with the incorporation of each nucleotide causing pyrophosphate (PPi) release. This stimulates ATP sulfurylase which is coupled to firefly luciferase, generating a flash of light which is in turn detected by a camera. This method of sequencing is known as pyrosequencing (Ronaghi *et al.* 1996).

The 454 Genome Sequencer was released in 2004 by Rothberg's company 454, and was the first commercially available NGS machine. It was later replaced by the more powerful 454 GS FLX. The company was bought by Roche in 2007 and the platform is often referred to as the "Roche 454".

The method can be separated into three separate stages; shotgun DNA library preparation (**Figure 1.7**), amplification of each of the library fragments by emulsion (**Figure 1.7**), and pyrosequencing (**Figure 1.8**).

This method of sequencing is capable of generating 400-600 million bases of sequence data per run. Sequencing runs take approximately 10 hours, with approximately two days of sample preparation required.

The increased sequence output and reduced time of emulsion PCR based clonal amplification and pyrosequencing compared to Sanger sequencing meant that for the first time sequencing projects could be undertaken that aimed to screen hundreds of genes at an economically viable cost. The first non-Sanger whole genome sequence of an individual human, James Watson, was sequenced using this method (Wheeler *et al.* 2008).

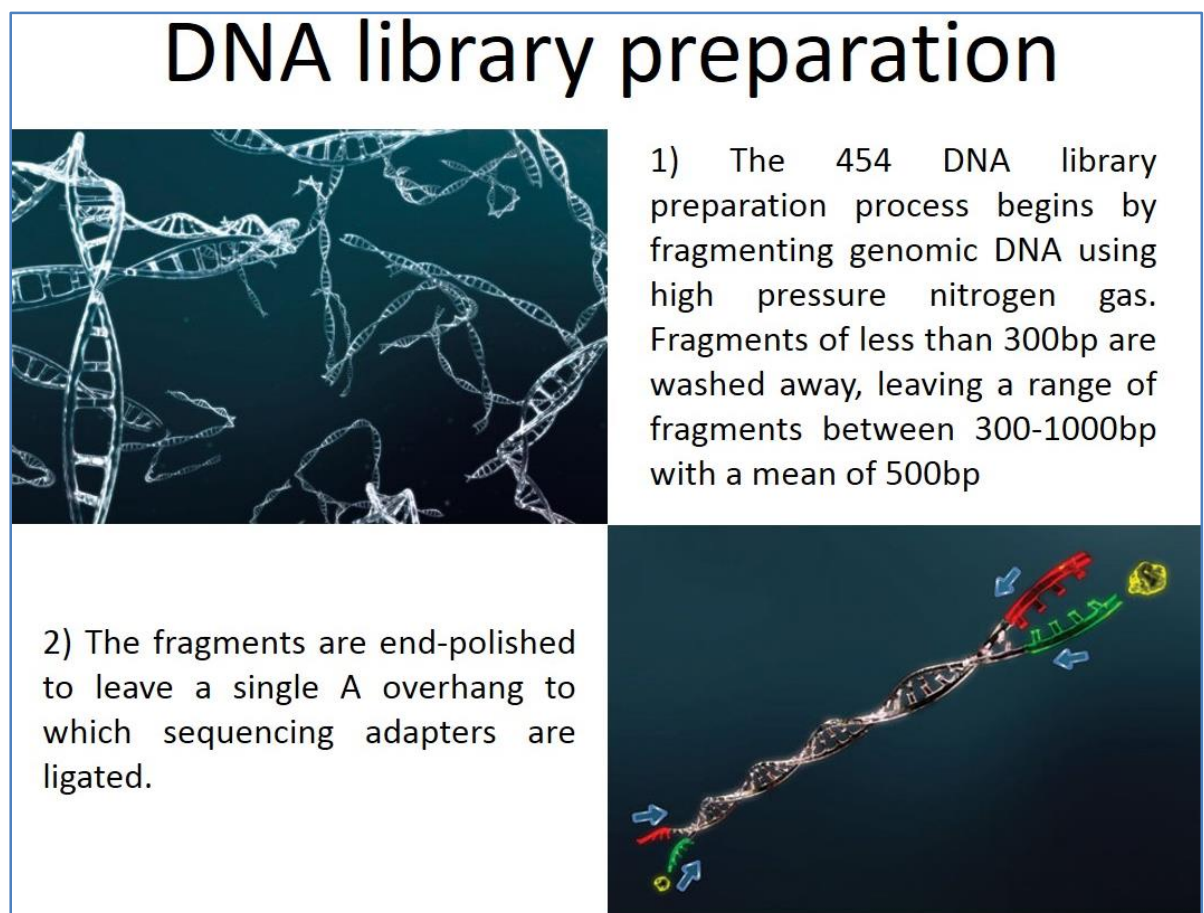
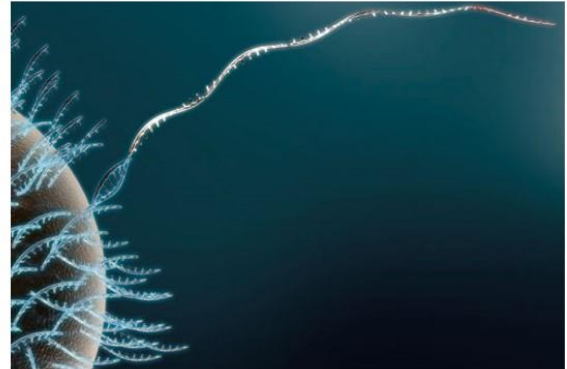


Figure 1.6 454 DNA Library preparation process. Genomic DNA is fragmented, end polished, and adaptors ligated to each end to create a DNA library.

Figure adapted from: <http://454.com/products/technology.asp>

Emulsion PCR

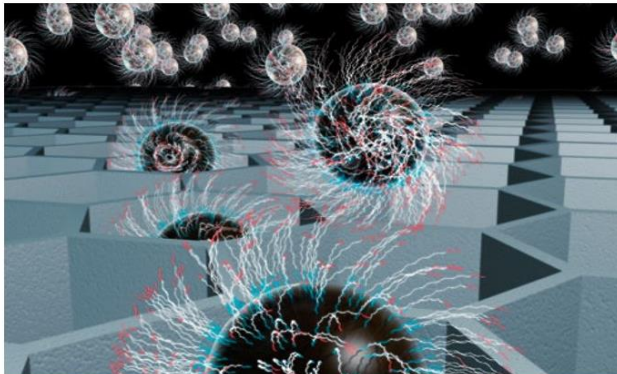
3) The fragments are attached to beads via the adapters. The concentration of the beads and DNA is calculated so that just one fragment attaches to each bead.



4) Each bead is emulsified in an oil-in-water emulsion creating independent microreactors. This allows clonal amplification of millions of fragments on each bead in individual PCR reactions without contamination of the amplicons.

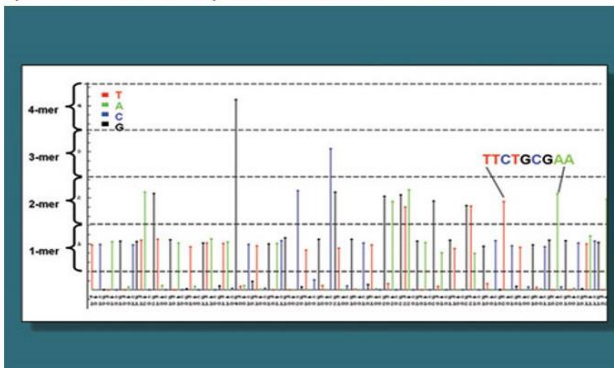
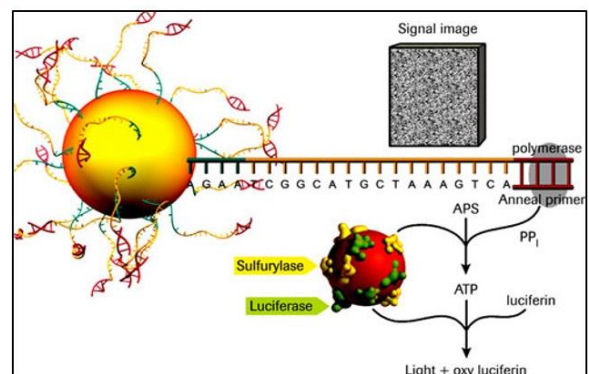
Figure 1.7 454 emulsion PCR process. DNA libraries are amplified by ligating them to microscopic beads, and emulsifying them in oil, creating independent microreactors in which PCR amplification can occur to create millions of fragments on each bead. The emulsion is then broken to allow the beads coated in clonally amplified fragments to be sequenced. Figure adapted from: <http://454.com/products/technology.asp>

Sequencing



5) At this point there will be hundreds of thousands of beads, each with millions of copies of a fragment of single stranded DNA attached to them. These beads are loaded onto a pico titer plate (PTP) which contains pico-litre scale wells that the beads are packed into.

6) The PTP is loaded onto the machine which bathes the device with DNA polymerase and primers. Nucleotides are added by the machine one at a time in a cycle. If the nucleotide currently bathing the device is incorporated onto the complementary strand then ATP is released. This causes a chemical reaction with luciferase that results in a flash of light, which is detected by a camera on the machine, indicating that the current nucleotide being added is present on that position of the strand.



7) As the machine cycles through the four nucleotides one by one, it is possible to determine which nucleotide is present by whether there is a flash. Homopolymer tracts are detectable by a longer flash, although this is not accurate for longer tracts and is one of the major disadvantages of pyrosequencing with luciferase. A high depth of coverage is therefore needed to ensure an accurate sequence is determined.

Figure 1.8 454 pyrosequencing. The beads coated in clonal fragments are loaded into a plate with microscopic wells, packed with the reagents required for synthesis of a complementary strand. The final reagent, the nucleotides, are incorporated in flows over the plate, and are then removed, with the machine cycling repeatedly through each nucleotide. When a nucleotide is incorporated into each position of the growing complementary strand pyrophosphate is released, which triggers a reaction coupled to the luciferase enzyme, which ultimately creates a flash of light which is detected by a camera on the machine. Homopolymer tracts are detected as longer, higher intensity flashes.

Figure adapted from: <http://454.com/products/technology.asp>

1.3.2 Sequencing by synthesis using reversible terminator chemistry

One of the most widely used methods of determining DNA sequences was developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge. It bears a much closer resemblance to Sanger sequencing than pyrosequencing, as it utilises dideoxynucleotides to terminate synthesis of a complementary DNA strand. However, the primary difference is that this chain termination is reversible, allowing synthesis of a complementary strand to be performed one nucleotide at a time (Bentley *et al.* 2008). Between the incorporation of each new nucleotide the identity of the dideoxynucleotide can be determined by fluorescent labelling and stimulation by a laser. This also allows homopolymer tracts to be identified accurately, preventing the detection problems encountered with some pyrosequencing methods.

As the chemistry used in this method incorporates only one nucleotide into the growing complementary strand at a time, it improves the accuracy of the sequence data generated but lengthens the run times to several days, as opposed to several hours with pyrosequencing, as the imaging step must take place in series rather than in parallel. However, the process can be scaled up to increase the throughput compared to pyrosequencing, and involves less hands-on time for preparation of samples.

The clonal amplification takes place on a solid surface called a flowcell in an automated process inside a machine called a cluster generator, rather than emulsion PCR amplification and emulsion breaks which have to be manually performed.

The newer sequencing machines which utilise this method allow two flowcells to be sequenced at the same time. Whilst one sample is being imaged by the machine, the other sample undergoes the chemistry to prepare it for the next round of imaging. This way the imaging process (the bottleneck in the sequencing stage) is performed continuously, increasing throughput by decreasing the time taken for sequencing runs. An overview of the entire methodology is shown in **Figure 1.9** and **Figure 1.10**.

Balasubramanian and Klenerman founded a start-up company called Solexa in 1998 using venture capital funding to commercialise the chemistry and create DNA sequencing machines. Solexa released its first sequencing machine, the Genome Analyzer in 2006 which was powerful enough to sequence one billion base pairs of data in a single run. Solexa was then bought by Illumina in 2007.

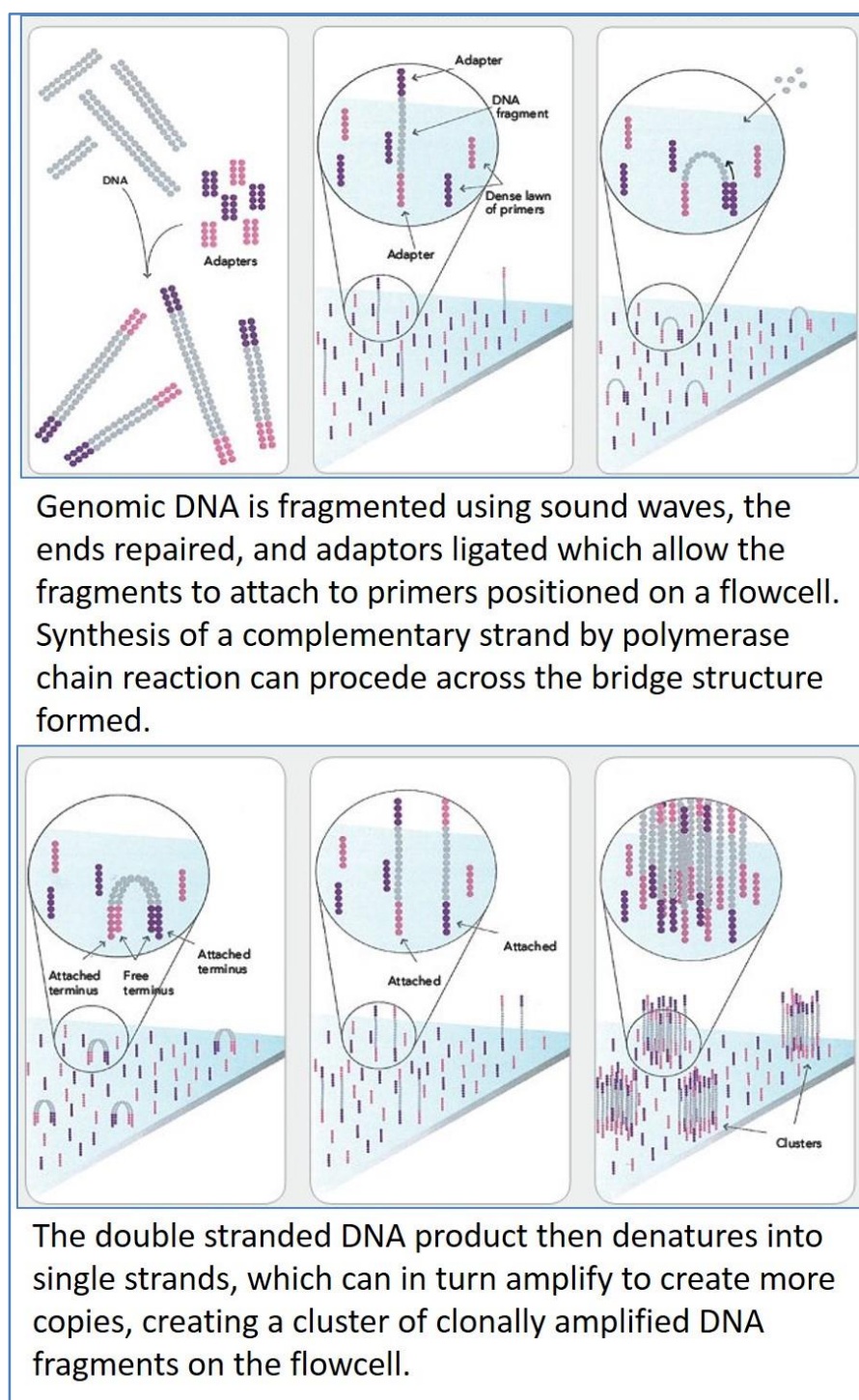
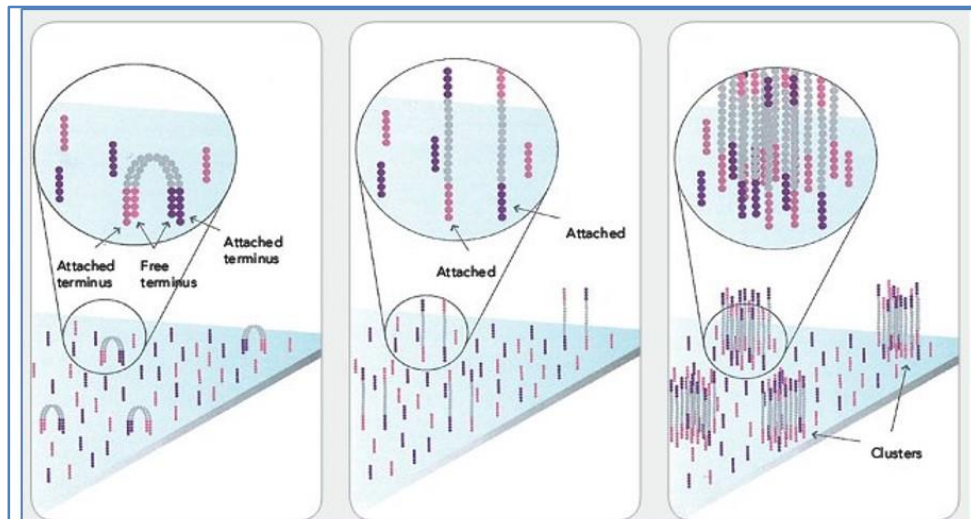
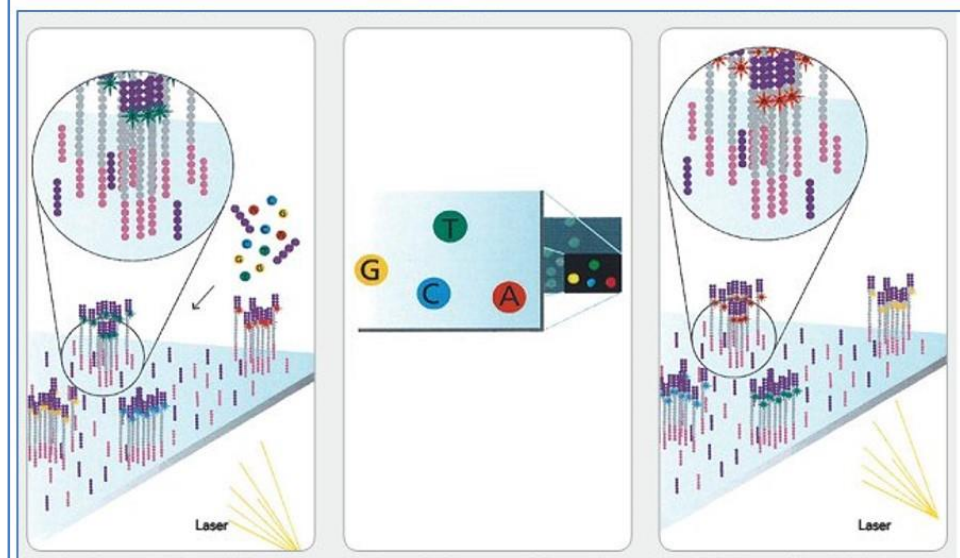


Figure 1.9 Illumina sequencing part 1.

Figure adapted from: <http://seqanswers.com/forums/showthread.php?t=21>



The double stranded DNA product then denatures into single strands, which can in turn amplify to create more copies, creating a cluster of clonally amplified DNA fragments on the flowcell.



After amplification of the cluster, the DNA is denatured. A complementary strand is then synthesised using dideoxy terminator nucleotides, which only allow the reaction to proceed one base at a time. These dideoxynucleotides are fluorescently labelled, allowing the identity of the nucleotide at each position to be determined by laser excitation. The dideoxynucleotide is then cleaved from the growing complementary strand, allowing it to extend one base further with each cycle.

Figure 1.10 Illumina sequencing part 2.

Figure taken from: <http://seqanswers.com/forums/showthread.php?t=21>

1.3.3 Whole genome sequencing using unchained base reads on self-assembling DNA nanoballs

A third NGS platform (used for investigations described in Chapter 6 of this thesis) was developed by Clifford Reid and Radoje Drmanac using unchained base reads on self-assembling DNA nanoballs (Drmanac *et al.* 2010).

The sequencing chemistry uses a novel library preparation method in which DNA is fragmented, and the fragmented molecules copied into a long continuous molecule consisting of many repeats of the fragment sequence. These molecules consolidate into “DNA nanoballs”; dense ball-like structures containing many repeated copies of the same sequence (**Figure 1.11 A**).

The DNA nanoballs are loaded onto silicon chips (**Figure 1.11 B**) where ligation based sequencing occurs using probes with one fluorescently labelled nucleotide that can identify the nucleotide being sequenced, surrounded by a random combination of nucleotides (**Figure 1.11 C**). Probes are manufactured with all possible combinations of nucleotides to allow binding to any position on the nanoball molecules (Drmanac *et al.* 1993).

Carnevali and colleagues (2012) reported the computational techniques required to assemble the data generated from the sequencing. Variants are called, annotated, and listed in a format called master variant (masterVar) files, which are a different format to the standard variant called files (VCFs) generated by most variant calling programs.

This method of sequencing by ligation was made available during the course of the investigations described in this thesis through a company called Complete Genomics, founded by Reid and Drmanac. Instead of providing sequencing machines to researchers, Complete Genomics perform all sequencing in-house on samples shipped to them.

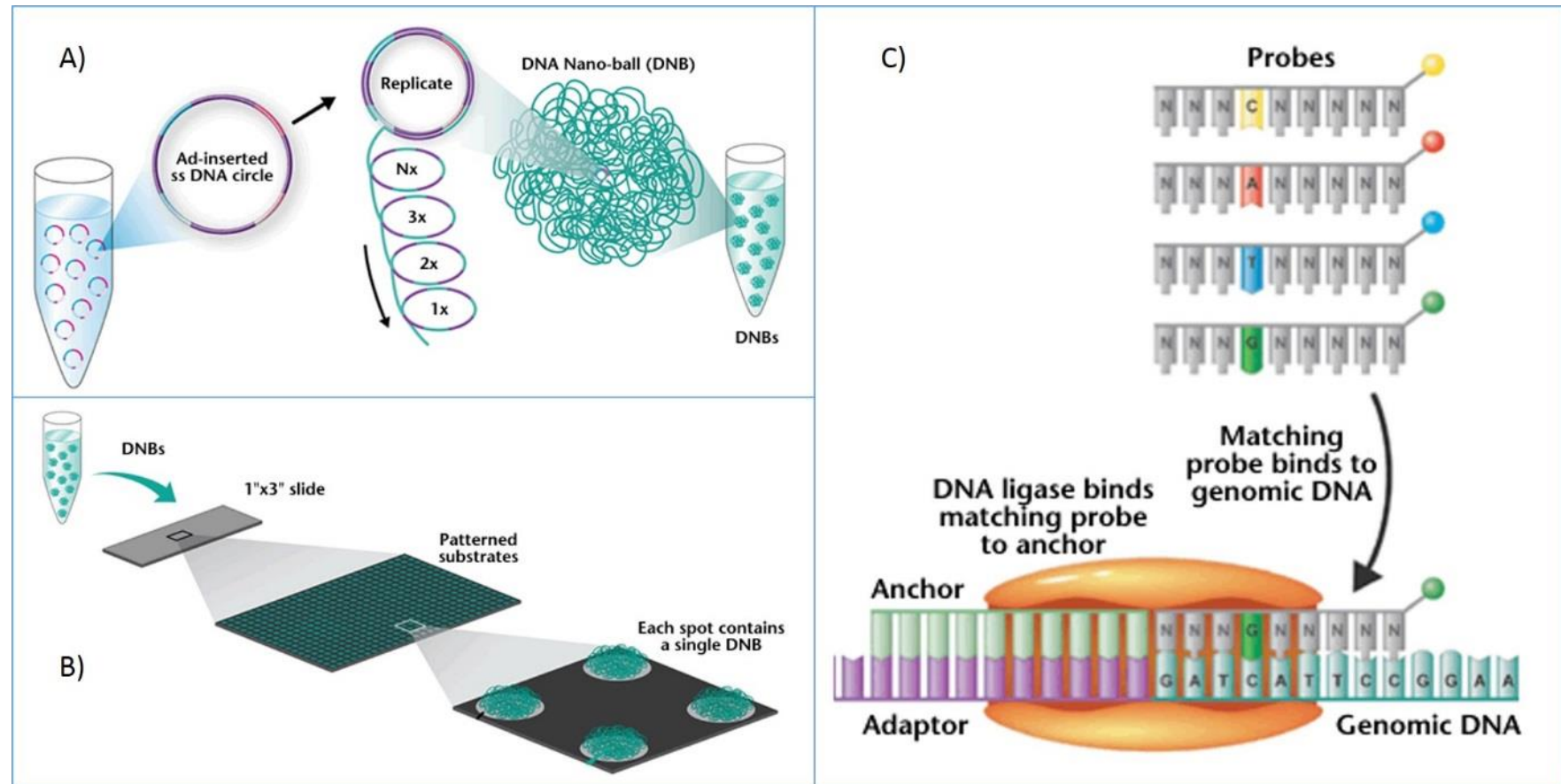


Figure 1.11 Complete Genomics sequencing technology. **A)** DNA is fragmented and copied to produce a long molecule containing many copies of the same fragment, which coil up into DNA nano-balls. **B)** DNA nanoballs are packed onto silicon chips. **C)** Ligation based sequencing determines the identity of nucleotides using fluorescently labelled probes. Figure adapted from: <http://www.completegenomics.com/technology/>

1.4 Sequencing costs and sequence capture

NGS technologies have dramatically reduced the cost of DNA sequencing as shown in **Figure 1.12**. From 2001 to mid-2007 the cost of genome sequencing approximately halved every two years due to gradual increases to the throughput of Sanger sequencing methods. The red arrow in **Figure 1.12** marks the point at which the cost of genome sequencing started to rapidly fall at a much faster rate due to the use of the first NGS technologies. However, at the time the work described in this thesis began in October 2010 the cost of sequencing a genome was still in the tens of thousands of dollars, which would be prohibitively expensive for use as a molecular diagnostic test.

NGS technologies are intended for high throughput sequencing. For many investigations including clinical diagnostics only a fraction of the entire genome is usually of interest: the coding sequences (referred to as the 'exome') and their flanking intronic sequence regulating splicing. Given the cost of sequencing each base it is unnecessary and very costly to sequence the entire genome. When specific disease associated genes are of interest, even exome sequencing can be unnecessarily wasteful. Therefore, techniques were developed to 'enrich' or 'capture' only those gene regions of interest for sequencing. These targeted approaches have been particularly beneficial in health care systems with very limited financial resources and where minimising time to diagnosis is important.

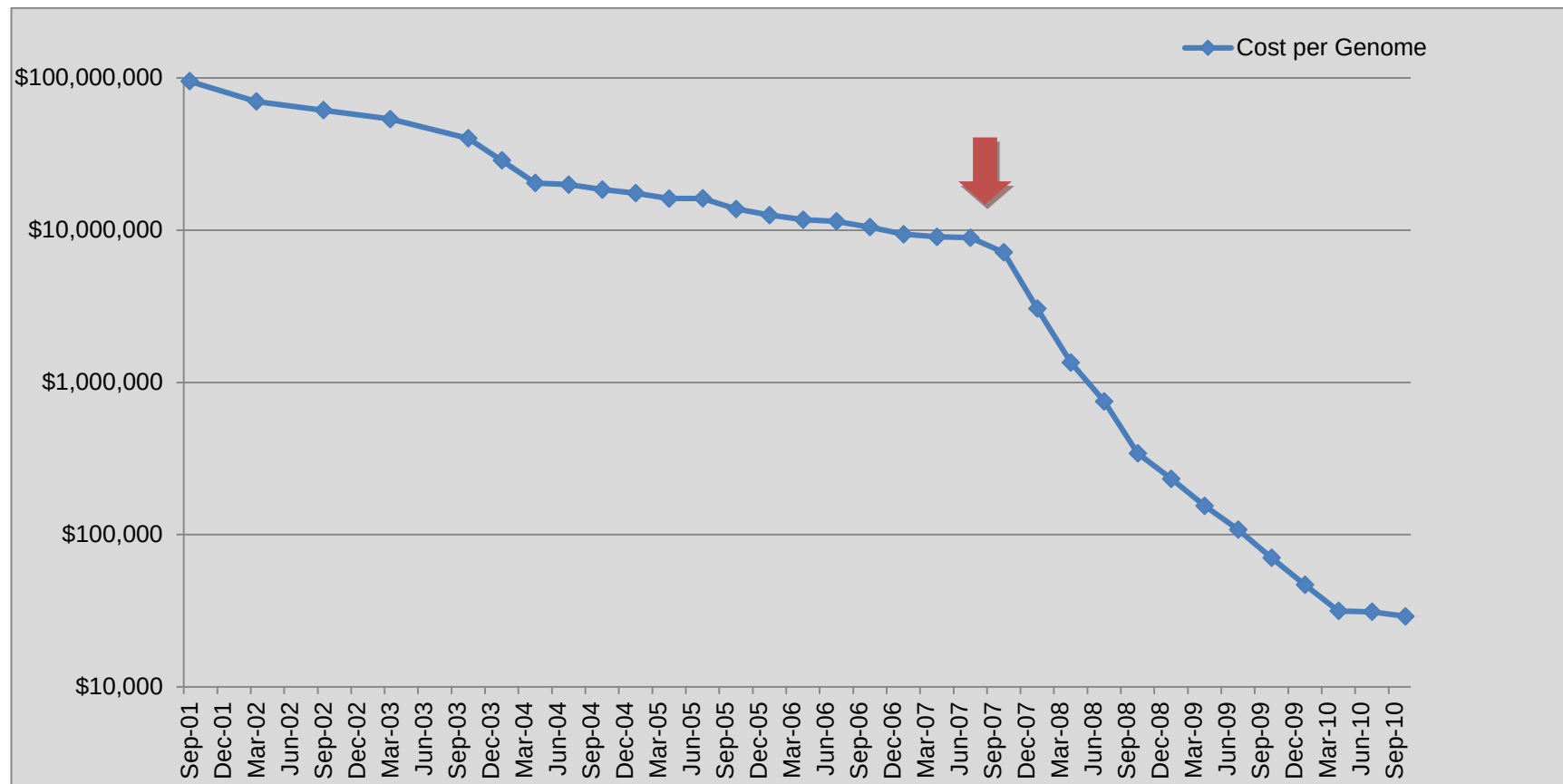


Figure 1.12 Cost of genome sequencing over time. The red arrow marks the point at which the first next-generation sequencing technologies became available for use. Data is only shown up until the point at which this project began, and was taken from: <http://www.genome.gov/sequencingcosts/>

1.5 DNA sequence capture and enrichment technologies

DNA sequence capture is an additional step in an NGS workflow, which can take place either before or after the preparation of a DNA library. It results in the removal of DNA fragments which do not contain regions of interest. The three methods used in this project are described below.

1.5.1 NimbleGen sequence capture array

The NimbleGen sequence capture array is a method of targeted DNA capture that utilises hybridisation of target sequences to a microarray (Albert *et al.* 2007). Amplified adaptor-ligated fragments of DNA libraries are hybridised to an array with oligonucleotide probes that have been designed to be complementary to the target regions of interest. DNA fragments that do not contain regions of interest are washed away. The hybridised target regions are then amplified by Ligation Mediated PCR (LM-PCR), a form of PCR where the primers bind to the adaptor sequence flanking the libraries. The captured and enriched DNA fragments are then used for downstream sequencing as normal (**Figure 1.13**).

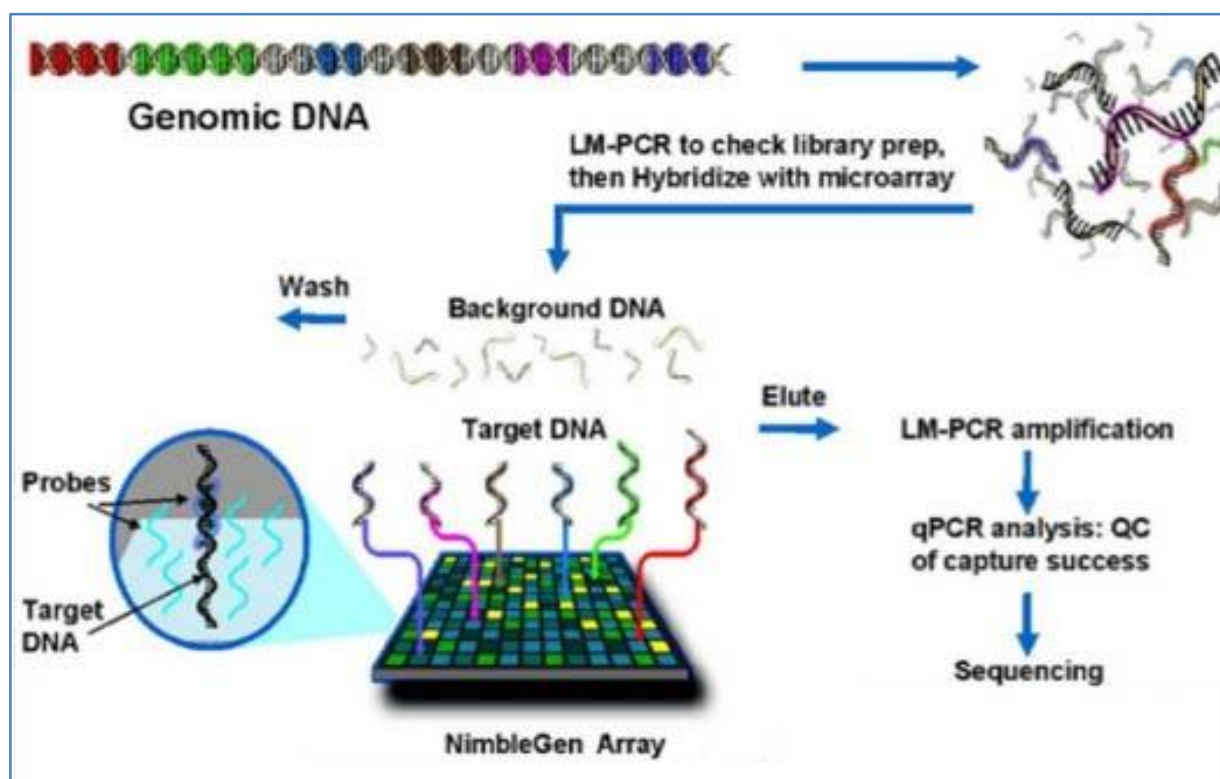


Figure 1.13 Schematic diagram of the NimbleGen sequence capture array workflow. Oligonucleotide fragments in a DNA library hybridise to probes on a solid surface array, capturing target sequence, and allowing non-target sequence to be washed away. Figure taken from: <http://www.mv163.cn/uploads/allimg/140312/1423431946-0.jpg>

1.5.2 Agilent SureSelect solution capture

Agilent SureSelect is also a hybridisation based capture method, but in contrast to the NimbleGen sequence capture array, SureSelect captures DNA in solution instead of on a solid surface, and uses 120-mer biotinylated cRNA baits rather than oligonucleotide probes (Gnirke *et al.* 2009). DNA which is complementary to the baits will hybridise to them, and the biotinylated baits then bind to streptavidin coated magnetic beads. A magnetic force applied to the tube containing the solution is then applied which attracts the magnetic beads with the bound bait and DNA of interest. The unbound DNA which is not of interest is washed away. The captured DNA is then resuspended, and the cRNA baits are digested, leaving the captured DNA fragments ready for sequencing (**Figure 1.14**).

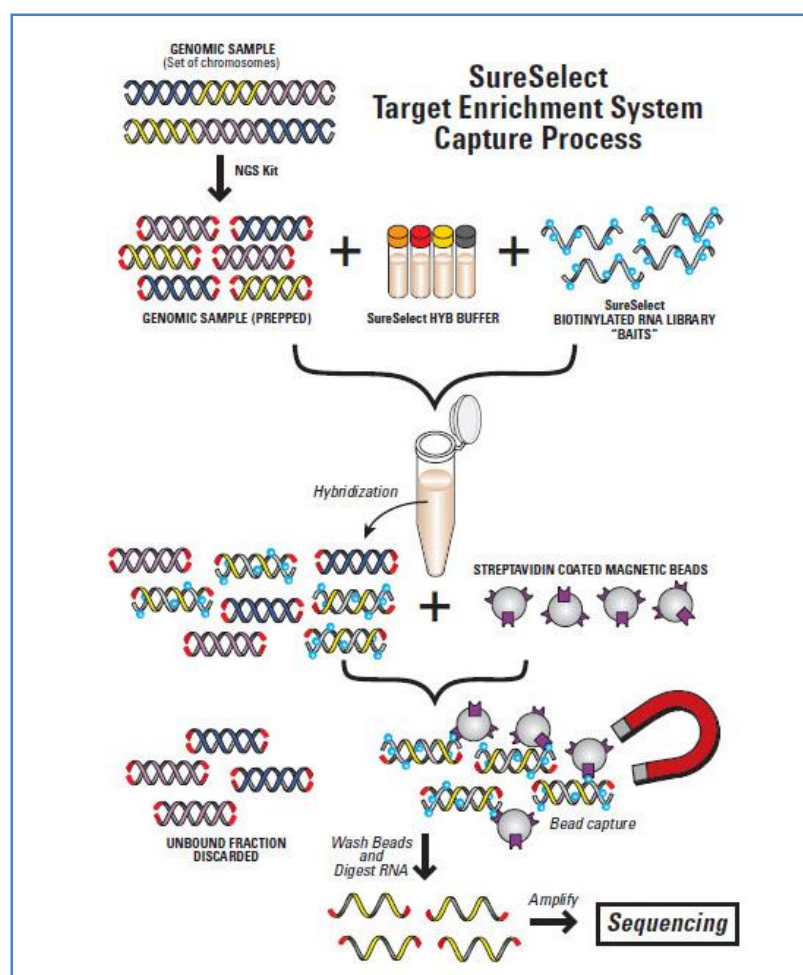


Figure 1.14 Schematic diagram of the SureSelect workflow. Target sequences of DNA fragments are captured by biotinylated RNA baits, which bind to streptavidin coated magnetic beads, allowing the target DNA to be pulled to one side of a tube while non-target sequence is washed away. The RNA baits are then digested to release the target DNA for sequencing. Image taken from: <http://www.genomics.agilent.com/>

1.5.3 RainDance microdroplet-based PCR

RainDance capture differs from the two previously described methods as it relies on PCR rather than hybridisation to amplify the targeted DNA of interest. It does this by using microfluidics and oil-in-water emulsion technology to allow reaction reagents and DNA to be independently generated, dispensed, merged, and for PCR to take place, all inside a single drop of water emulsified by oil (Brouzes *et al.* 2009) as shown in **Figure 1.15**.

The genomic DNA is dispensed into individual droplets of water, as are the primer pairs for each amplification, and the two are merged together (**Figure 1.15**). A Polymerase Chain Reaction then occurs inside the droplet, amplifying the specific DNA sequence of the primer design, which can then be sequenced on an NGS platform.

All three of these sequence capture and enrichment methods were tested during this thesis and the results comparing their advantages and disadvantages are presented in chapter 3.

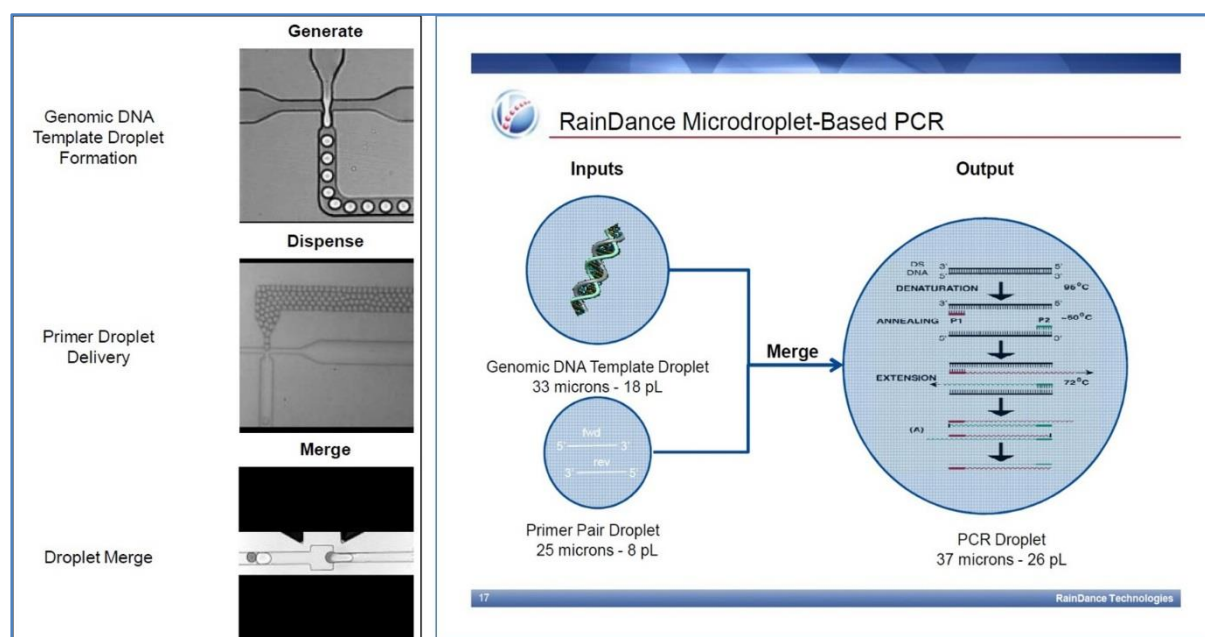


Figure 1.15 RainDance microfluidics technology. Left: Images of the different types of microfluidic processes performed in RainDance. Right: Schematic showing how genomic DNA droplets and primer droplets can be merged to create independent PCR micro reactors Images adapted from: <http://www.RainDancetech.com/>

1.6 Bioinformatic analysis of sequencing data

Sequencing data is initially produced in a binary format and requires conversion during the analysis pipeline. Each of the vast number of sequence reads has a unique identifier and quality scores for the calling of each base in the read. Files containing this raw data are called SFF for the 454, and SAM for Illumina reads.

Sequence assembly software such as “Stampy” (Lunter and Goodson 2011, used as a personal communication before publication) is used to assemble the reads into a continuous sequence, by mapping each read to the most current build of the human genome reference sequence. Overlapping reads can be overlaid on top of one another to build up a continuous sequence (Figure 1.16).

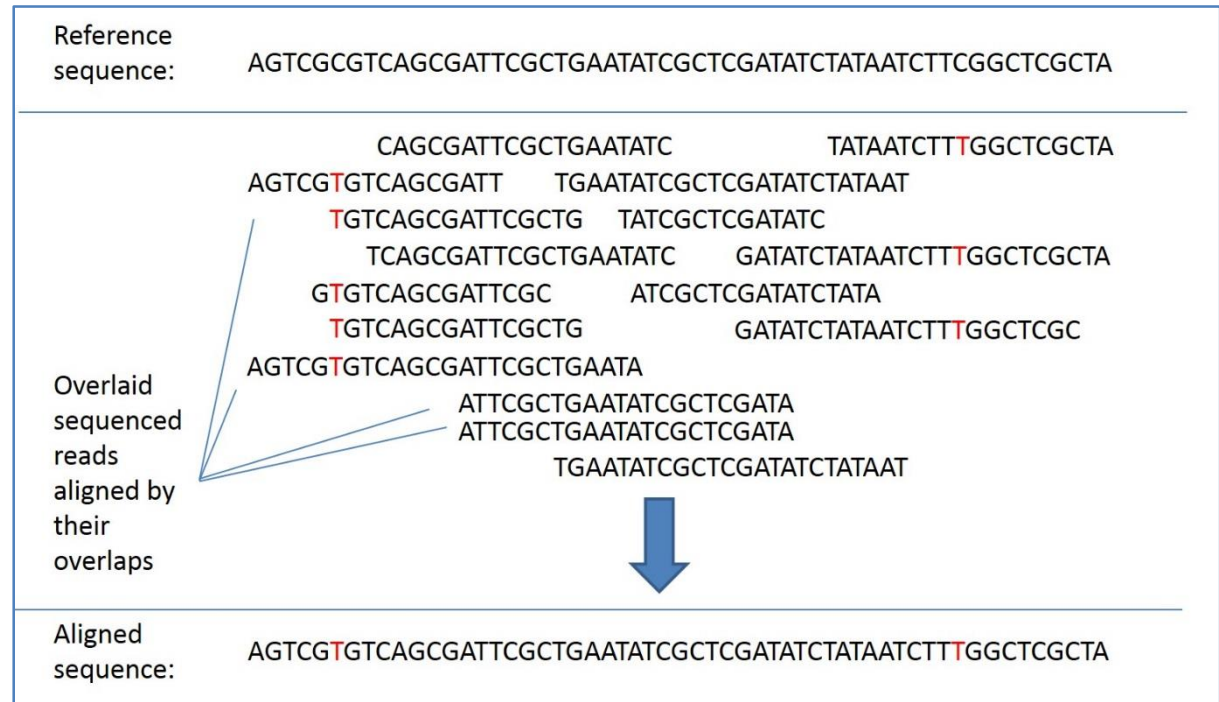


Figure 1.16 Alignment of NGS reads. Reads generated by NGS are aligned by overlaying them where they overlap using the reference sequence, to create a continuous consensus sequence. Variants can be detected by identifying bases which differ from the reference sequence, indicated here in red.

1.6.1 Evaluating alignment

Sequence capture and enrichment technologies are designed to selectively amplify DNA in regions of interest. The target region can vary from a single exon to an entire exome. An enrichment technology with 100% efficiency amplifies only target regions of DNA, and removes all other DNA prior to sequencing, so that every read sequenced maps to the target region, and no reads would map to any other region of the genome.

Since no methods are 100% efficient, it is important to determine the percentage of sequence reads mapping back to the target region. This is an indication of the efficiency of capture and is termed alignment. A higher alignment makes it more likely that the regions of interest will be covered adequately and makes sequencing more cost effective.

1.6.2 Coverage of bases

The consensus sequence assembled from NGS data is made up of multiple reads. The number of reads covering a particular base of the sequence is known as the depth of coverage for that base (**Figure 1.17**). A minimum depth of coverage is needed to ensure the sequence is accurate, which can be mathematically determined by taking into account the length of the genome, the number of reads, the length of the sequence reads, and the amount of sequence needed to detect an overlap (Sims *et al.* 2014).

A higher level of coverage leads to a higher statistical probability that the sequence is correct, and makes it more likely the sequence will correctly identify heterozygous alleles which can be incorrectly called as homozygous at low coverage depths. One of the first studies of reversible terminator chemistry data by Bentley and colleagues showed that the accuracy of heterozygous variant calls increased significantly as coverage rose, up to a depth of approximately 30x coverage (Bentley *et al.* 2008).

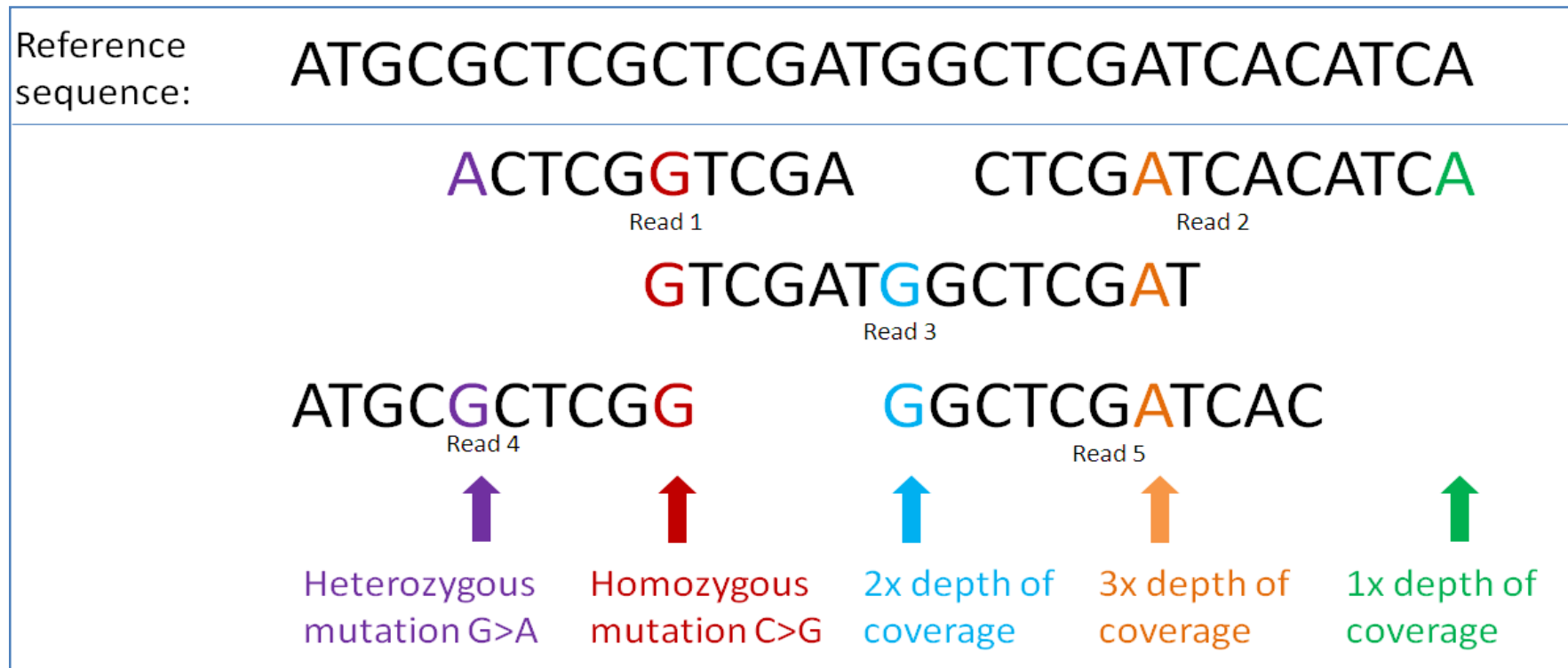


Figure 1.17 Simplified example of how shotgun sequencing data is compiled *in silico*. Reads are aligned to overlapping sequences of a reference sequence. Depth of coverage refers to the number of reads covering a certain nucleotide. Variants can be called by identifying changes between the reference sequence and the assembled sequence, with heterozygous variants expected to appear in approximately half of the reads, and homozygous variants in all of them.

1.6.3 Variant calling and analysis

The DNA of any individual will show natural genetic variation from the human genome reference sequence, which will include SNPs and other polymorphisms (see section 1.1 p. 19), as well as potentially pathogenic mutations. A list of such variants is compiled from next-generation sequencing data by variant calling software such as “Platypus” (Rimmer *et al.* 2014, used as a personal communication before publication).

Variant calling software can identify sequence variants by comparing the sample of interest with a reference sequence, which in most cases will be the most current version of the human genome reference sequence. The reference sequence is based on the sequences of anonymous individuals from a variety of ethnic groups who were selected by an ethics panel at the start of the human genome project (Venter *et al.* 2001). In projects searching for mutations causing a Mendelian disorder, most variants identified will be non-pathogenic and must be filtered out to determine either the single or pair of variants that cause the disease.

Variant called files (VCFs) that are produced by variant calling programs will contain only basic information on the change in sequence detected by the sequencing chemistry used, which might include some data that can be used for quality control, such as the frequency of reads in which the variant appears. The initial stages of variant analysis therefore require these variants to be annotated with additional information about the location of the variant within a gene (e.g. exonic/coding, intronic, untranslated region, intergenic, etc.). Further annotation can then be added to each variant, such as the effect on amino acid substitution for a SNV in a coding region, or the predicted effect on splicing for variants present in essential splice sites. This can then be further annotated to include predicted functional effects using *in silico* prediction programs.

1.6.3.1 *In silico* methods for predicting pathogenicity

Mutations causing Mendelian disorders often have highly significant effects on protein function, such as producing stop codons or frameshift mutations. These are often easily recognisable and software scripts designed to prioritise these variants are straightforward. However, missense

mutations are produced by single nucleotide variants (SNVs) which can be difficult to distinguish from the many benign single nucleotide polymorphisms (SNPs) that will be present in any sequence data. A variety of pathogenicity prediction programs have been developed to analyse such missense variants for characteristics which might suggest pathogenicity, such as the level of conservation of that nucleotide among related species, the type of amino acid change caused, and the position of this amino acid in the protein relative to functionally important domains.

A commonly used pathogenicity prediction program is PolyPhen2 (Polymorphism Phenotyping Version 2). PolyPhen2 predicts the effects of single nucleotide variants (SNVs) by use of eight sequence-based and three structure-based predictive features that compare the wild-type protein to the mutant protein (Adzhubei *et al.* 2010). Pathogenicity scores are determined using database information on multiple-sequence alignment which shows how highly conserved the nucleotide is at that position (Adzhubei *et al.* 2010). Nucleotides that are highly conserved by evolution across multiple species are likely to code for amino acids that have a functionally significant role in the protein, so changes here are more likely to be pathogenic. The prediction features also take into account whether a variant allele is present in a hypermutable site where variations in the DNA sequence are likely to be insignificant to the functioning of the protein (Adzhubei *et al.* 2010).

Another commonly used program for pathogenicity prediction is SIFT (Sorting Intolerant From Tolerant). SIFT assumes that important positions in a protein sequence may have been conserved throughout evolution and therefore substitutions at these positions may affect protein function. Thus, by using sequence homology, SIFT predicts the effect of all possible substitutions at each position in the protein sequence (Kumar *et al.* 2009).

Other commonly used programs include MutationTaster (Schwarz *et al.* 2010) which has a similar function to PolyPhen2; Human Splicing Finder (Desmet *et al.* 2009), and GeneSplicer (Pertea *et al.* 2001), all of which predict the effect on splice site changes; and ESEFinder (Smith *et al.* 2006) and RESCUE-ESE (Fairbrother *et al.* 2002) which predict the effect of changes to

exonic splicing enhancer sequences. Other approaches include the use of curated databases of mutations that have been shown in a publication to cause pathogenicity, such as HGMD Professional (www.hgmd.org).

Ultimately *in silico* methods can only give an approximate indicator as to how important a specific nucleotide is to the functioning of the protein it codes for, and cannot be relied on to give a firm molecular diagnosis. To confirm with certainty that a variant is pathogenic, a difference from wild type must be demonstrated, usually using a cellular model or animal assay.

In addition, a crucial method of determining disease causality will be to show segregation of the disease with the mutation, by demonstrating that all those affected in a family have the same genotype compared to those who are healthy. However, where there is reduced penetrance, genetic heterogeneity, or phenocopies, segregation may not be helpful.

1.6.3.2 Segregation analysis

Analysing the segregation of variants between diseased individuals and unaffected members of the same family can be a powerful method for determining the pathogenicity of a variant. In recessive conditions a pathogenic mutation should be identified on both alleles of an affected individual. These may be either homozygous or compound heterozygous variants. Unaffected family members may also be carriers of the disease, harbouring a single pathogenic mutation. Usually a single pathogenic variant is found in each unaffected carrier parent of a patient. In some cases both mutations may not be identified and there are numerous explanations for this, which are discussed in detail in section 6.1.4 p. 209.

In dominant disorders only one pathogenic variant is needed to cause the disease phenotype. Such conditions are more likely to be caused by *de novo* mutations compared to recessive conditions, as there is a stronger selection pressure operating against the variants, which cannot be spread by carriers. If a condition is known to be dominant then it is easier to filter out benign variants, as any variant which is present in unaffected family members can be excluded.

1.6.3.3 Determining allele frequencies in diseased and control populations

As well as analysing the segregation of a putative pathogenic variant in affected and unaffected family members, the frequency of a variant in databases containing large populations of control and disease chromosomes can be used to give an indication of the probability of pathogenicity.

DNA sequence variants that directly cause a rare genetic disorder should appear at a low frequency in control populations, so with a large enough data set it should be possible to preclude many variants from being pathogenic if they appear at a high percentage in control populations.

The need to perform such an analysis led to the creation of the 1000 genomes project (<http://www.1000genomes.org>), which aimed to sequence the genomes of 1000 individuals across four ethnic groups to develop a comprehensive map of human genetic variation. At the time this project commenced this resource was the largest database of genome sequences available from which data on allele frequencies could be obtained, so the allele frequencies in the bioinformatic analysis pipeline in Chapter 4 uses data from this source.

In the time since these investigations began however, further databases of exome and genome sequences have been produced and made available, which include the exome variant server (<http://evs.gs.washington.edu/EVS>) and ExAC (<http://exac.broadinstitute.org>).

Subsequent to the submission and examination of this thesis, allele frequencies were obtained for variants identified in this thesis from the exome variant server (EVS), and have been added at the relevant points in the results chapters.

1.6.4 Creating a bioinformatic pipeline

The previous sections have illustrated the different analyses that are required for data generated using NGS platforms. To do this efficiently bioinformatic pipelines must be developed that allow consistent results to be produced, and which can be adjusted with experience and to suit different analysis requirements. At the time of starting these investigations such pipelines were in the very early stages of development, and Chapter 4 describes the development of one.

Figure 1.18 shows what is required of such a pipeline; sequence alignment of the raw reads generated by the sequencing machine, variant calling, variant annotation, and variant filtering. In most cases a separate computer program is used to perform each of these processes, each of which will produce a different type of file containing the newly processed data.

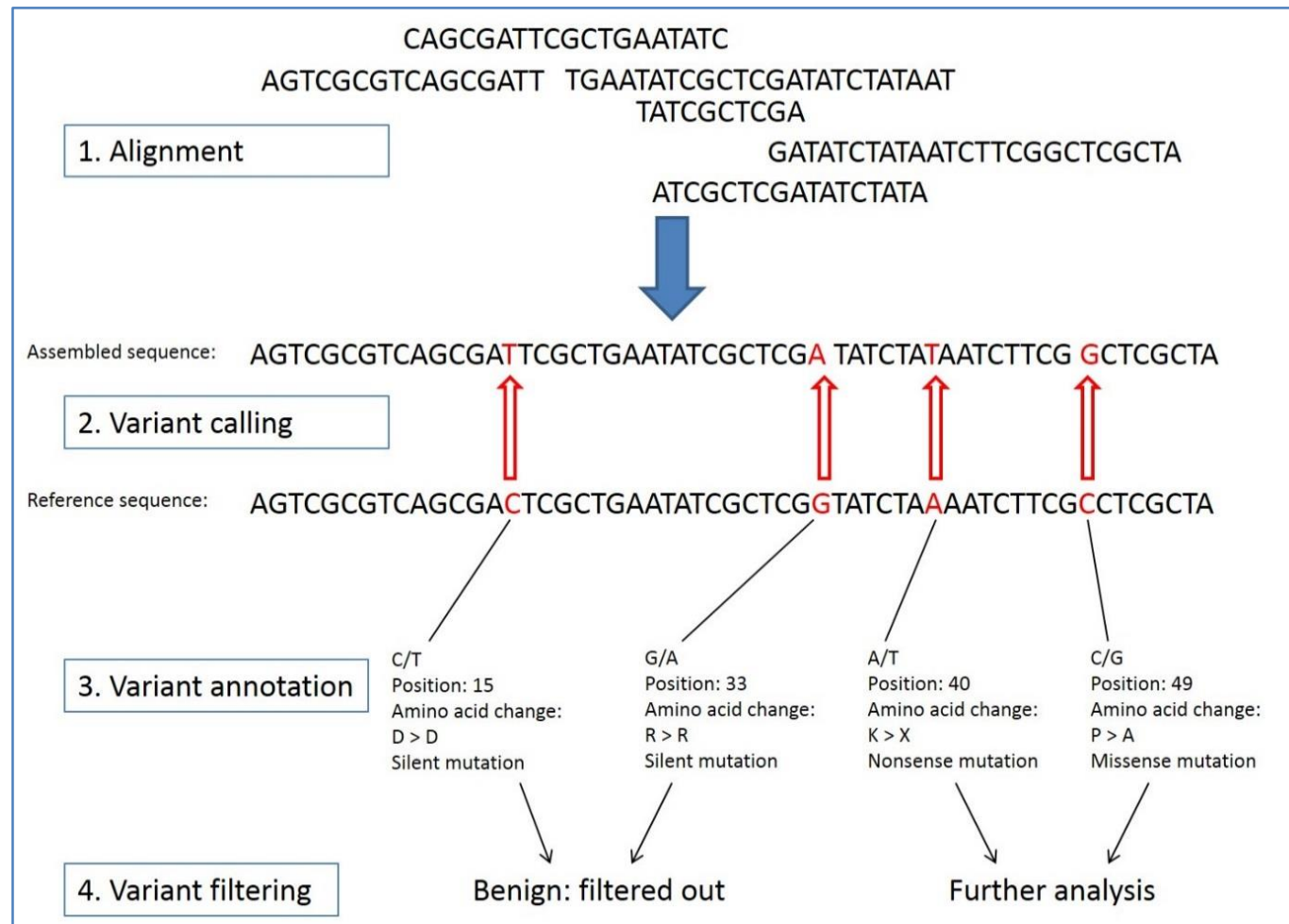


Figure 1.18 Bioinformatic pipeline. (1) Raw data generated by sequence runs is aligned to create an assembled sequence. (2) This sequence is compared to a reference sequence to call variants. (3) The variants are annotated. (4) Variants that have no coding effect are filtered out, while nonsense mutations and missense mutations can be further analysed with pathogenicity prediction programs and other tests.

1.7 Clinical utility of genetic testing

Since the first discovery of mutations in DNA being associated with human disease (Ingram 1956) there have been major advances in diagnostics for genetic disorders. Important advantages of making such a diagnosis include:

- Confirming a clinical diagnosis.
- Identifying inheritance patterns for genetic counselling including reproductive options such as prenatal testing or pre-implantation genetic diagnosis (Sohrab *et al.* 2010).
- In some cases, providing prognostic information.

Furthermore, this can be used to:

- Identify novel genes causing disease.
- Identify novel molecular mechanisms of disease.
- Develop novel treatment strategies based on genetic mechanism or understanding of the molecular pathways of the disease.
- Identify genetic modifiers for Mendelian traits. These may influence the expression of the condition and may also be amenable to treatment strategies.
- Develop pharmacogenetic treatments based on knowledge of the mutant proteins.

1.7.1 Overview of genetic technologies used in diagnostics

Genetic diagnostics relied on a variety of different techniques before the introduction of NGS:

- The early experiments in genetics after the rediscovery of Mendel's work led to the discovery of genetic linkage (Morgan 1911), the notion that traits are often inherited together because the genes are located in close physical proximity on the same chromosome. Linkage studies are still commonly performed and allow the physical location of genes to be mapped on to chromosomes, often by using SNPs as markers.
- Karyotyping was one of the earliest techniques used to determine changes to DNA structure. The technique involves visually observing chromosomes during metaphase or

prometaphase to observe any differences or anomalies in the size or shape of chromosomes, centromeres, satellites, or heterochromatic regions. This is usually achieved by halting cell division using colchicine and using dyes to stain the chromosomes, allowing them to be visually observed using light microscopy.

- Southern blots can be used to detect changes in the size of DNA fragments. The technique was developed by Edwin Southern and involves separation of DNA fragments by electrophoresis, followed by transfer to a filter membrane, and detection of separated fragments by probe hybridisation (Southern 1975).
- Sanger sequencing, introduced in section 1.2.1 p. 24, is another technique which can accurately detect changes in the sequence of DNA (Sanger *et al.* 1977).
- Restriction fragment length polymorphism is a technique which uses restriction digest enzymes to cleave DNA at specific positions. The fragments generated are separated by agarose gel electrophoresis and transferred to nitrocellulose paper in a Southern blot. The fragments are then detected by use of complementary DNA probes which hybridise to fragments of the same length. Differences in the size of the alleles detected can be used to give information on linkage of genetic loci, so this technique was used for many years to map genes (Saiki *et al.* 1985).
- Polymerase chain reaction (PCR) can be used to detect changes in the size of DNA fragments when combined with a method such as agarose gel electrophoresis (Mullis and Faloona 1987).
- Single strand conformation polymorphism (SSCP) is a process in which DNA is denatured into single strands. These strands are then separated by agarose gel electrophoresis. Single stranded DNA coils up on itself in a conformation that depends on sequence. The structures of these conformations provide resistance to the movement of the strands, allowing sequence changes between different strands to be detected by measuring and comparing the distance each strand has moved (Orita *et al.* 1989).

- Heteroduplex analysis is a procedure similar to SSCP, in which PCR products are denatured by heating to 95°C, then slowly allowed to cool, causing heteroduplexes to form where there is a DNA mismatch. These heteroduplexes influence the rate of migration of DNA bands during agarose gel electrophoresis, allowing detection of single base changes in a sequence (White *et al.* 1992).
- DNA Microarrays (sometimes known as DNA chips) are a high-throughput evolution of the DNA detection principles first used in the Southern blot. They are microfabricated arrays of oligonucleotide probes which can hybridise to DNA samples, allowing genotyping of the target sequence (Maskos and Southern 1992).
- Array comparative genomic hybridisation (arrayCGH) can detect DNA sequence copy number variations by comparing a test sample to a reference sample. The samples are hybridised to an array of oligonucleotide probes which fluoresce either yellow to signify equal copy numbers between the samples, green if there are higher levels in the test sample, or red if there are lower levels in the test sample (Kallioniemi *et al.* 1992).
- Arrayed primer extension (APEX) microarrays also make use of microarray hybridisation technology, however the hybridisation is performed on fragmented PCR products. The hybridised sequence is then extended one nucleotide further using a fluorescently labelled terminator nucleotide, which can identify the base at that position (Kurg *et al.* 2000, Tonisson *et al.* 2000).
- Multiplex ligation probe amplification (MLPA) involves the ligation of probes to a target sequence of DNA followed by amplification of the probes using PCR. Only when both of a pair of probes bind to a target sequence and ligate to one another can the amplification proceed, which allows the detection of copy number variants and other changes to sequence (Schouten *et al.* 2002).

All of these techniques have been used to provide information that has a clinical utility in the diagnosis and treatment of genetic diseases, however their throughput limits the amount of information that can be obtained from them.

1.8 Genetic disorders studied in this project

The majority of the work in this project is focussed on inherited retinal dystrophies, with some experiments described in chapter five performed on samples from families with Ohtahara syndrome, an early onset epilepsy syndrome.

1.8.1 Inherited retinal dystrophies (IRD)

Inherited retinal dystrophies (IRD) are a group of disorders that cause visual loss and impairment due to abnormal function of the retina, and in some cases associated structures. Many of the conditions are degenerative, with evidence of loss of photoreceptors or other retinal cell types, but in some cases the conditions are not associated with degeneration, for example congenital stationary night blindness.

The retina is a part of the central nervous system located in the eye (**Figure 1.19 A**) which comprises multiple layers of cells. The eye enables vision by focussing light that enters through the pupil onto the retina where it is detected by two different types of photoreceptor cells, known as rods and cones (**Figure 1.19 B**). In humans cone cells are mostly located in a dense cluster near the centre of the retina, known as the fovea. There are three different types of cone cell that can detect different wavelengths of light, allowing colour vision. The fovea, located within the macula is also responsible for fine vision, due to the density of photoreceptors contained in this region of the retina.

Rod cells are found in the whole of the retina except for the foveola, the central part of the fovea where cone cells are at their densest. Rod cells are activated in lower levels of light intensity and are therefore the primary cells subserving peripheral and night vision.

Light is detected by rhodopsin in rod cells, and by cone opsins in cone cells. Rhodopsin is encoded by the *RHO* gene, and the structure of the protein is shown in **Figure 1.20**. The N terminus lies on the extracellular surface of the protein, and the C terminus on the intracellular side. There are seven transmembrane segments numbered H1-H7, and an amphipathic α -helix stretch labelled H8 running perpendicular to them (Menon *et al.* 2001).

Mutations in the *RHO* gene were first shown to cause retinitis pigmentosa by Dryja *et al.* (1990) who went on to report further mutations in the gene as being a major cause of RP (Dryja *et al.* 2000). By inducing mutations in the gene in regions coding for the different domains of the protein, Franke *et al.* (Franke *et al.* 1990) were able to show that alterations to the cytoplasmic loops did not affect binding of 11-*cis* retinal, but prevented stimulation of the guanosine triphosphatase activity of the G protein transducin (G_t), indicating that at least the second and third loops are required for correct functioning of Rhodopsin, and so mutations that affect these regions can cause a disease phenotype (Franke *et al.* 1990).

In addition to Rhodopsin humans have four other opsin molecules: the three cone opsins (I, II, and III) which are used for colour vision, and melanopsin which is present in photosensitive retinal ganglion cells that control circadian rhythm, and other responses of the body to changes in light intensity.

The stimulation of an opsin molecule triggers the visual phototransduction cascade (shown in **Figure 1.21** for rod cells). The stimulation of rhodopsin triggers photoisomerisation of 11-*cis*-retinal to all-*trans*-retinal. The further steps of the cycle involve reduction, oxidation, and esterification steps which convert the substrates back and forth between all-*trans*-retinal, all-*trans*-retinol, retinyl esters, 11-*cis*-retinol, and 11-*cis*-retinal. Many of the genes coding for proteins that are involved in the phototransduction cascade have been found to have mutations which cause retinal diseases (reviewed in Lamb and Pugh 2006). Some of these genes are also shown in **Figure 1.21**.

The key areas of the retina can be seen in the fundus image in **Figure 1.22**. The retina is a key structure, and mutations in different genes which are expressed in the retina can lead to disorders of vision which may have different clinical appearances and hence a wide range of different phenotypes, collectively known as inherited retinal dystrophies (IRD). There are various different terms used to describe these phenotypes, but they can broadly be divided into three

main types: those that primarily affect the peripheral/mid-peripheral vision but may involve the centre later in disease, those that primarily affect the centre of vision, and those that affect both.

Other terms used to describe inherited retinal dystrophies include eponymous classifications which refer to the clinician who first described the condition, such as Stargardt disease. Eponymous terms are often still used but are now usually supplemented by other pertinent clinical or functional features including phenotypic appearance, functional performance (including electrophysiological recordings and psychophysical measurements), and in some cases genotype.

Diseases that primarily affect the peripheral or mid peripheral vision have historically been labelled as Retinitis Pigmentosa (RP), a specific term originally coined by Dutch ophthalmologist Franciscus Donders (Donders 1884). It is classically described as a retinal dystrophy characterised by intra-retinal bone spicule pigmentation, pale waxy discs and attenuated vessels, which usually presents as loss of both night vision (nyctalopia) and peripheral vision leading to 'tunnel vision' and eventually blindness (Fairbrother *et al.* 2002). This particular type of IRD has a primary deterioration of rod cells; usually the centre is involved at late stages in the disease. The image on the left of **Figure 1.23** shows extensive retinal loss with heavy intra-retinal pigment and preservation of the central retina. The image on the right of **Figure 1.23** shows typical intra-retinal bone spicule pigmentation in the inferior retina with attenuation of vessels.

In contrast macular dystrophies are characterised by loss of central vision, inability to see detail, abnormal colour vision and sometimes delayed dark adaptation. Macular diseases may have both rod and cone involvement. If limited in extent or relatively benign the photoreceptor loss may be restricted to the macula, and not affect global cone or rod function. Images of a type of macular dystrophy known as Stargardt disease are shown in **Figure 1.24**.

There are some conditions in which both cones and rods are affected, resulting in both central and peripheral vision loss at a similar time in the disease. Cone-rod dystrophies may show

temporal differences in cell populations being affected. Examples of such cone-rod dystrophies are shown in **Figure 1.25** and **Figure 1.26**.

1.8.1.1 The genetic basis of inherited retinal disorders

The different phenotypes that are observed in these disorders can be traced to the genes that are mutated and the function of the proteins they encode. The mutation of the *RHO* gene in RP has already been introduced, but genes that code for other structures in the retina can give rise to a range of different phenotypes.

Mutations in the retinal pigment epithelium 65 kDa (*RPE65*) gene can cause Leber congenital amaurosis (Marlhens *et al.* 1997), a type of severe, early-onset retinal dystrophy which affects both cones and rods, and also retinitis pigmentosa (Gu *et al.* 1997). The protein encoded by the gene was shown to be the source of the isomerohydrolase activity which converts all-trans retinyl ester to 11-cis retinol in the retinal pigment epithelium (Moiseyev *et al.* 2005).

A variety of genes in which mutations can cause RP have been named or renamed after the condition. Mutations in the *RP1*, *RP2* and *RP9* all cause retinitis pigmentosa through different inheritance patterns. One questionnaire placed the percentage of each inheritance pattern for RP at 84% autosomal recessive, 10% autosomal dominant, and 6% autosomal dominant (Boughman *et al.* 1980).

Mutations in the *RP1* gene were first associated with causing RP in an autosomal dominant inheritance pattern (Pierce *et al.* 1999). An extensive screening of 56 families with autosomal dominant RP (adRP) by Bowne *et al.* (1999) showed that mutations appeared to cluster in a 443 nucleotide segment of exon 4 of the *RP1* gene. They estimated that of the approximately 1 in 3500 people worldwide affected by RP, in approximately 7% of these patients the cause of the condition is a mutation in *RP1* (Bowne *et al.* 1999). Later work showed an association between mutations in the gene and autosomal recessive RP (Khaliq *et al.* 2005). The *RP1* gene was originally identified in mice as having an association with retinal hypoxia, which suggested its function may be involved in retinal neovascularisation, however later work showed that its role may be involved in cilia regulation (Liu *et al.* 2002).

Mutations in the *RP2* gene cause X-linked retinitis pigmentosa (Schwahn *et al.* 1998). Some work has been performed to attempt to elucidate the mechanism by which the protein produced by mutated forms of the gene cause disease. Studies of mutant and wild type *RP2* proteins have suggested that acyl modification occurs in the N-terminals by myristoylation and palmitoylation, and that mutations which result in these modifications not occurring can therefore affect the trafficking of the protein within the cell by preventing targeting of the *RP2* to the plasma membrane (Chapple *et al.* 2000).

While mutations in *RP1* are suspected to have an involvement in ciliary function, and mutations in *RP2* have an X-linked inheritance pattern, mutations in the retinitis pigmentosa GTPase regulator-interacting protein (*RPGRIP1*) gene have been shown to have both. Mutations in this gene cause Leber congenital amaurosis (Dryja *et al.* 2001) and cone-rod dystrophy (Hameed *et al.* 2003). Functional analysis of the protein showed that it associates *in vivo* with RanBP2 at the nuclear rim in a transformed murine photoreceptor line, and suggested that a change in the interaction between these molecules can be the cause of the disease (Castagnet *et al.* 2003).

Other mechanisms that can cause a disease phenotype include changes to the function of a transporter which can prevent the clearing of toxic compounds from a photoreceptor, ultimately resulting in the death of the cell. Such a defect is the cause of the various *ABCA4* retinopathies, which are introduced in more detail in section 6.1 p.194.

Finally, mutations in homeobox genes such as the cone-rod homeobox (*CRX*) gene can cause a cone-rod dystrophy phenotype by preventing the differentiation of cilia cell types into photoreceptors (Akagi *et al.* 2004).

The range of mechanisms by which a genetic mutation can result in the development of a disease phenotype is vast, and only a few have been covered here. When analysing variants identified by next-generation sequencing it is important to investigate the function of a gene and investigate the potential mechanisms by which mutations in the gene can result in a functional effect. In addition, the inheritance patterns of the disease should match the known inheritance patterns that have been identified, and the incidence rate should also be in line with what would be expected from either a recessive, dominant, or X-linked condition.

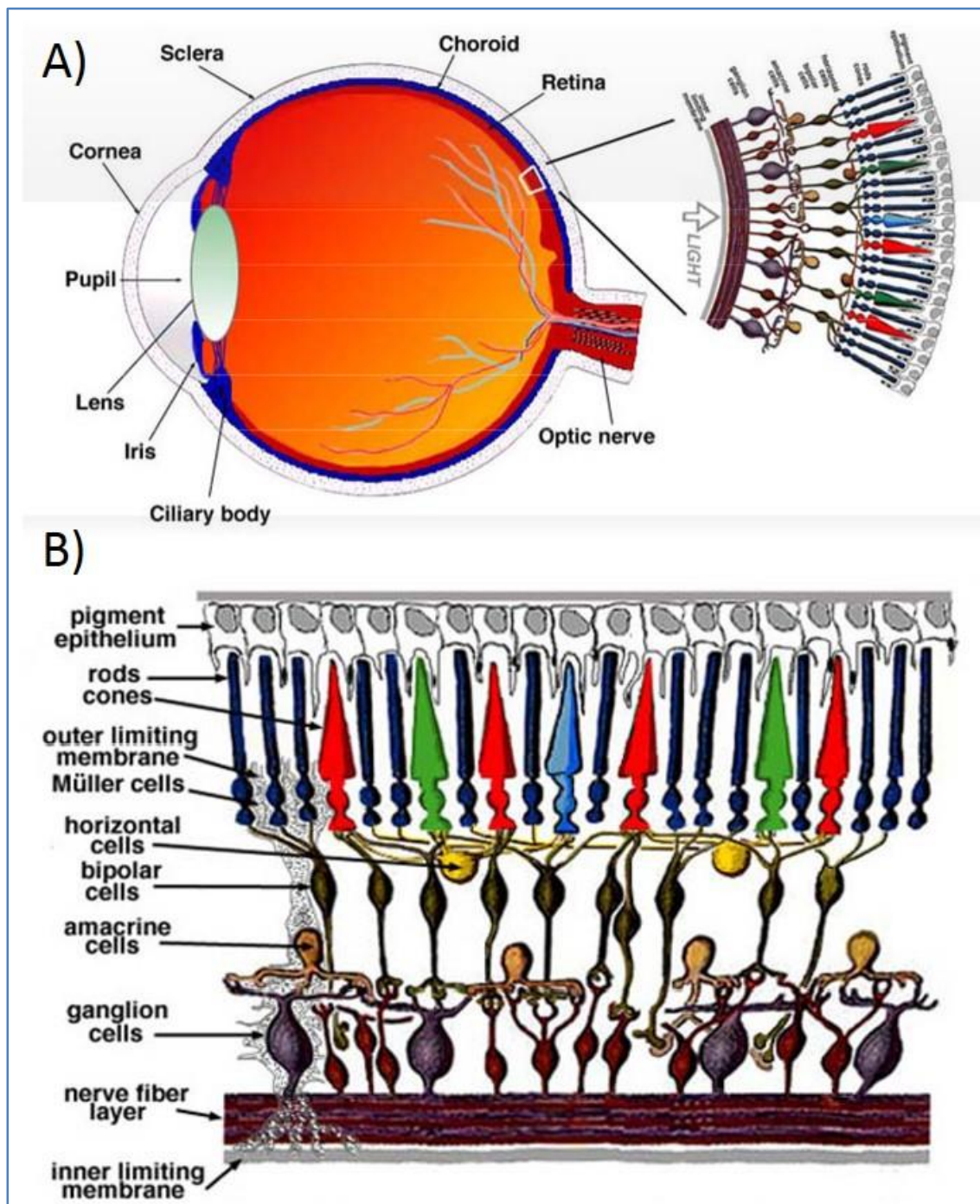


Figure 1.19 The mammalian eye. **A)** Schematic section of a human eye with a schematic enlargement of the retina. Light enters the eye through an aperture formed by the cornea, iris and lens, which focus the light onto the neural retina. The retina covers the rear inner globe of the eye and neuronal signals are transmitted from the retina to the retinorecipient areas of the brain via the optic tract. **B)** Schematic diagram of the six major cell types and their organisation. The rods and cones convey visual information to the ganglion cells via the second order bipolar cells. At the outer plexiform layer, horizontal cells facilitate lateral connectivity and feedback to the photoreceptors. At the inner plexiform layer amacrine cells allow lateral connection between bipolar and ganglion cells. The optic nerve is formed by the axons of all ganglion cells. A Muller cell is also shown. Image taken from <http://webvision.med.utah.edu/>

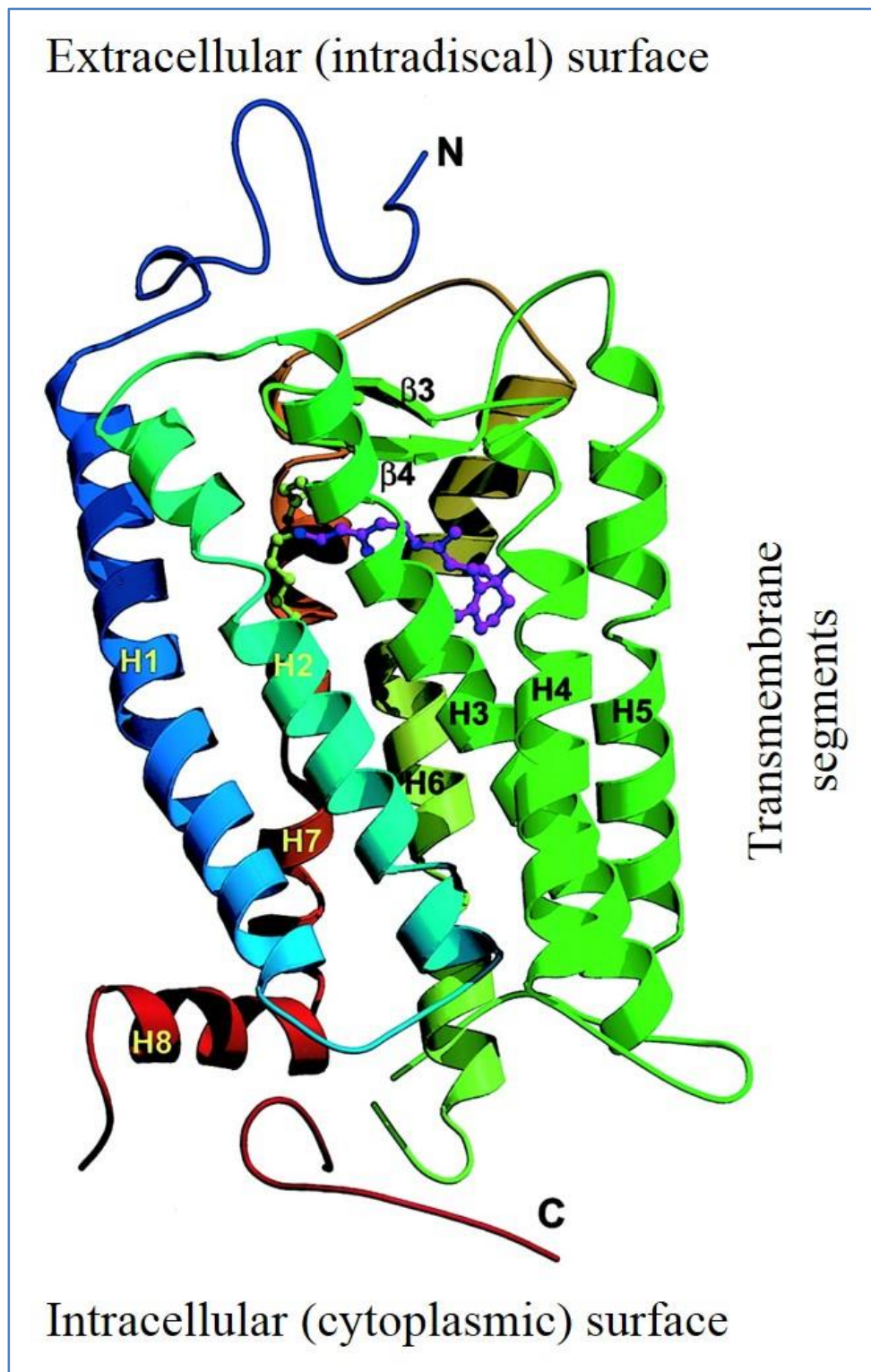
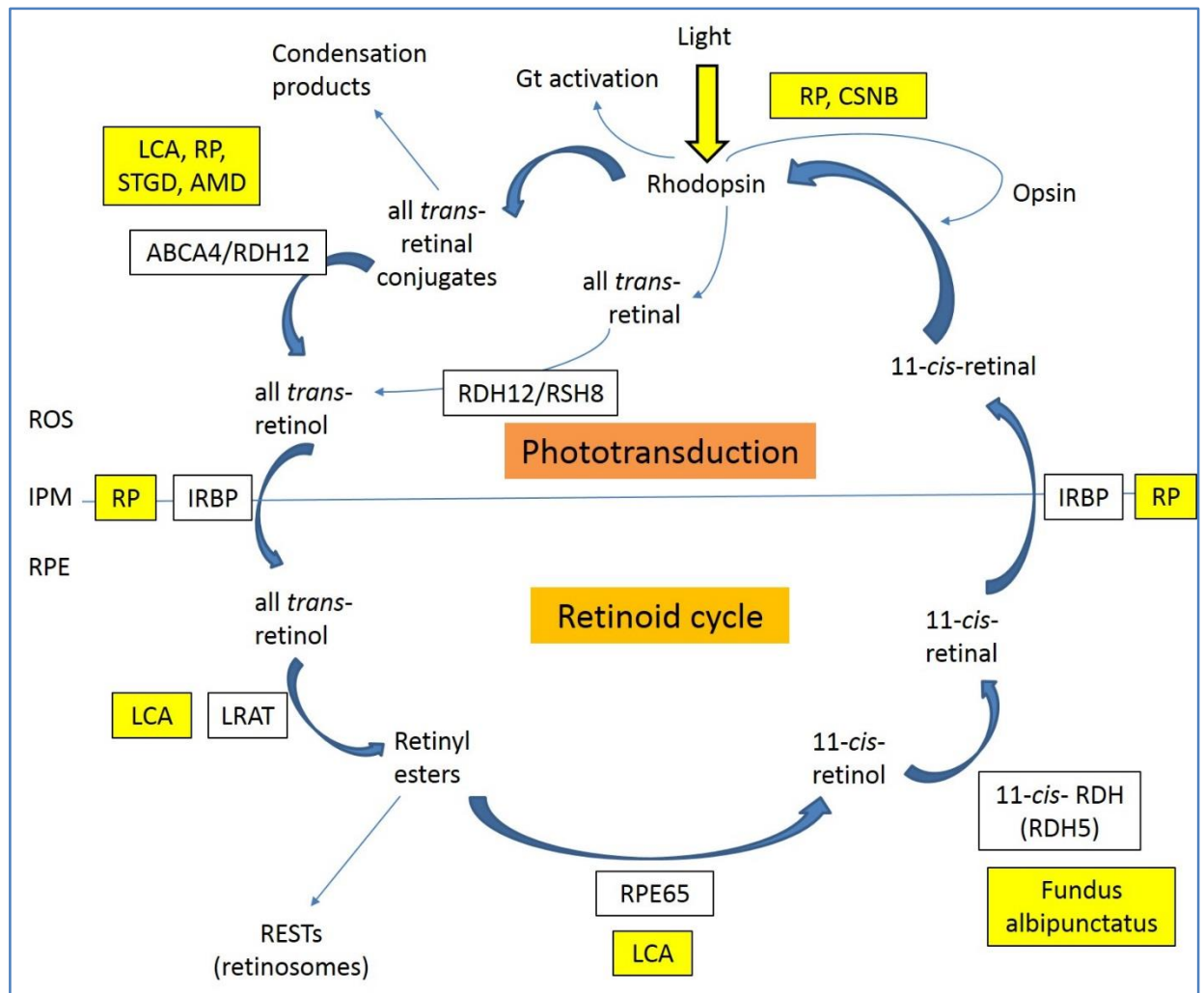


Figure 1.20 A molecular graphics ribbon diagram of the structure of the rhodopsin protein. The N terminus lies on the extracellular surface of the protein, and the C terminus on the intracellular side. There are seven transmembrane segments numbered H1-H7, and an amphipathic α -helix stretch labelled H8 running perpendicular to them. Part of the structure joining H8 to the C terminus has not been resolved and is not shown. Figure adapted from Menon *et al.* (2001).



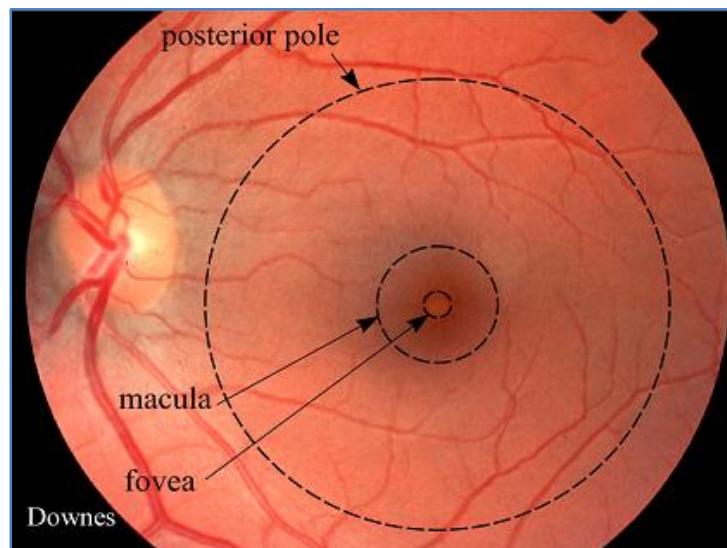


Figure 1.22 Fundus image of a normal retina. The macula contains a high proportion of cone cells, which are most densely clustered in the fovea. The posterior pole marks the area of the retina between the optic disc (the bright yellow oval on the left of the image) and the macula (Image provided by S. Downes).

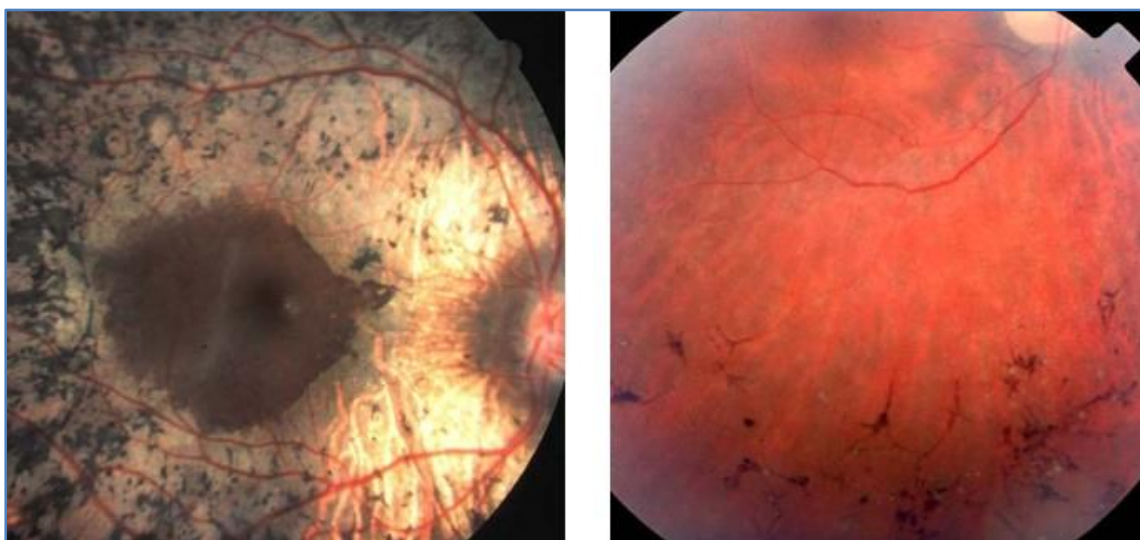


Figure 1.23 Retinal images of retinitis pigmentosa. Images provided by A. Bird and S. Downes (personal communication).

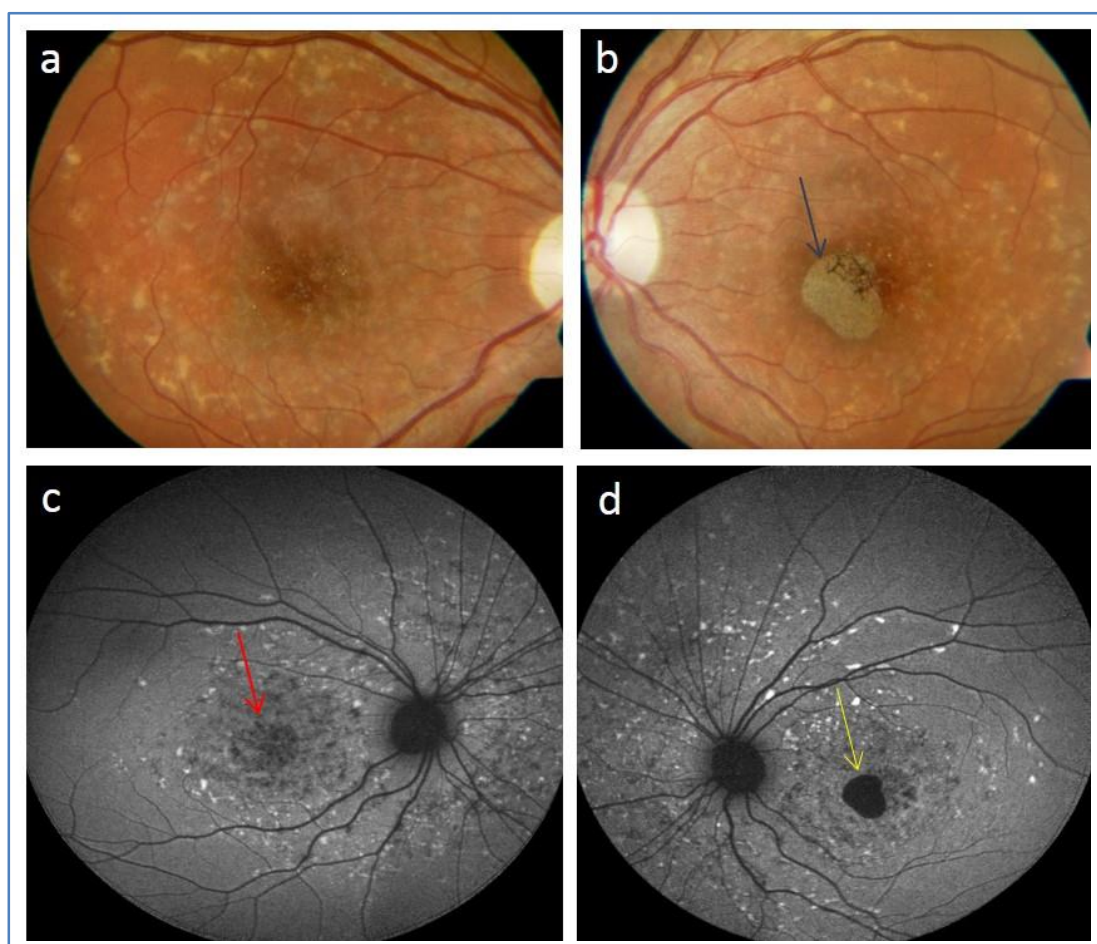


Figure 1.24 Images of Stargardt disease. a) Colour images of Stargardt disease showing multiple flecks with minimal central atrophy. b) Colour image of the other eye showing similar fleck deposition but also significant central atrophy (see blue arrow), c) Autofluorescence imaging of the eye in figure a, showing early atrophy (reduced signal – see red arrow), d) Autofluorescence imaging of the eye in figure b showing marked central atrophy (reduced signal – see yellow arrow). Images provided by S. Downes (personal communication).

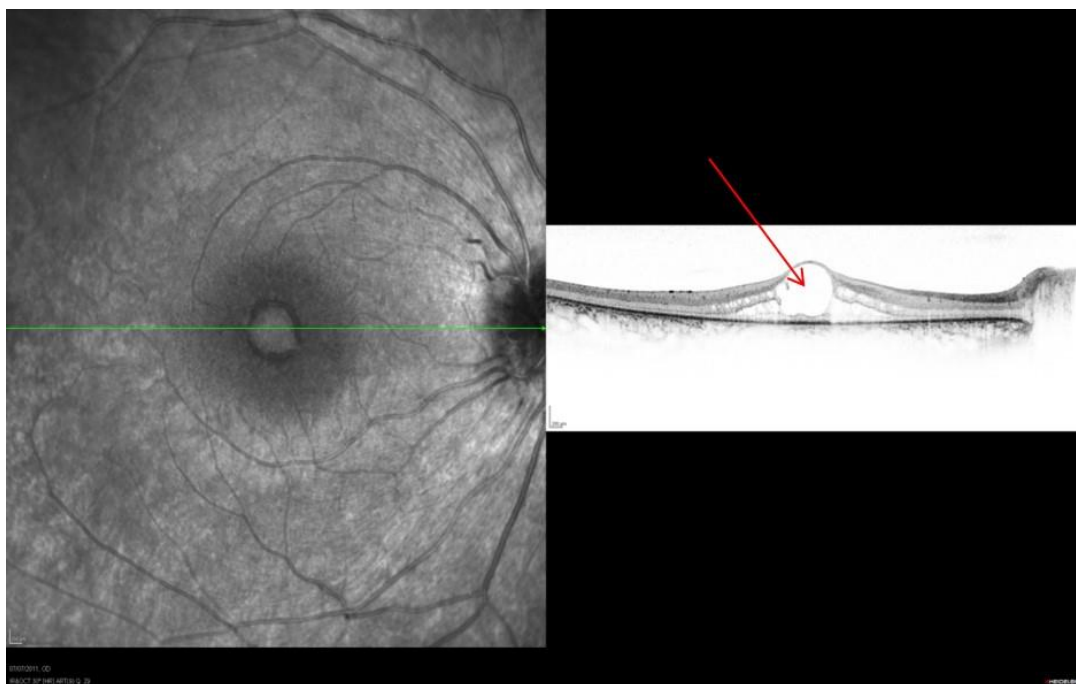


Figure 1.25 Imaging of patient with cone-rod dystrophy. who has cystoid macular oedema as demonstrated by optical coherence tomography imaging showing the central cystic spaces (red arrow). Images provided by S. Downes.

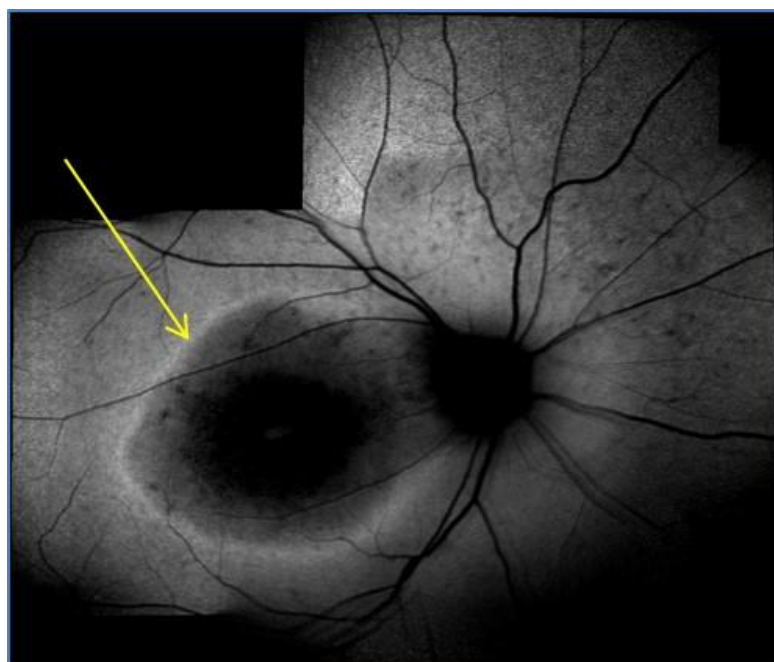


Figure 1.26 Autofluorescence image of the right eye of a patient with cone-rod dystrophy. showing typical annulus of increased signal (yellow arrow) within arcades seen in some IRD, with decreased signal centrally demonstrating atrophy. Image provided by S. Downes.

1.8.2 Genetic testing for retinal disorders

At the time the work described in this thesis began in October 2010 over 150 genes had been identified as having an association with retinal disease (Daiger *et al.* 1995) as shown in **Figure 1.27**. These genes are involved in a range of functions, from structural elements, to the creation of light sensitive pigments, to the transport of molecules across membranes. Mutations in these genes can therefore be associated with a wide range of clinical phenotypes. The clinical and genetic heterogeneity of IRDs makes assigning an accurate diagnosis based on clinical data alone very challenging.

Before the emergence of NGS, genetic testing for retinal disorders was limited to testing for a few genes, often on a research or academic basis using the techniques described in section 1.7.1 p. 51.

These methods placed a restriction on the numbers of genes that could be interrogated to determine the genetic cause of a patient's condition, due to cost and time constraints. Clinicians use the appearance and natural history of the condition as seen in the patient as well as any family history details to try to determine the inheritance pattern of a disease, and then request appropriate tests. This is very restrictive as not all potentially relevant genes or mutations could be tested and the test direction is heavily dependent on the acumen and experience of the clinician.

One of the primary aims of this work is therefore to evaluate the ability of NGS technologies to be used as a clinical diagnostic screen for a large number of IRD genes.

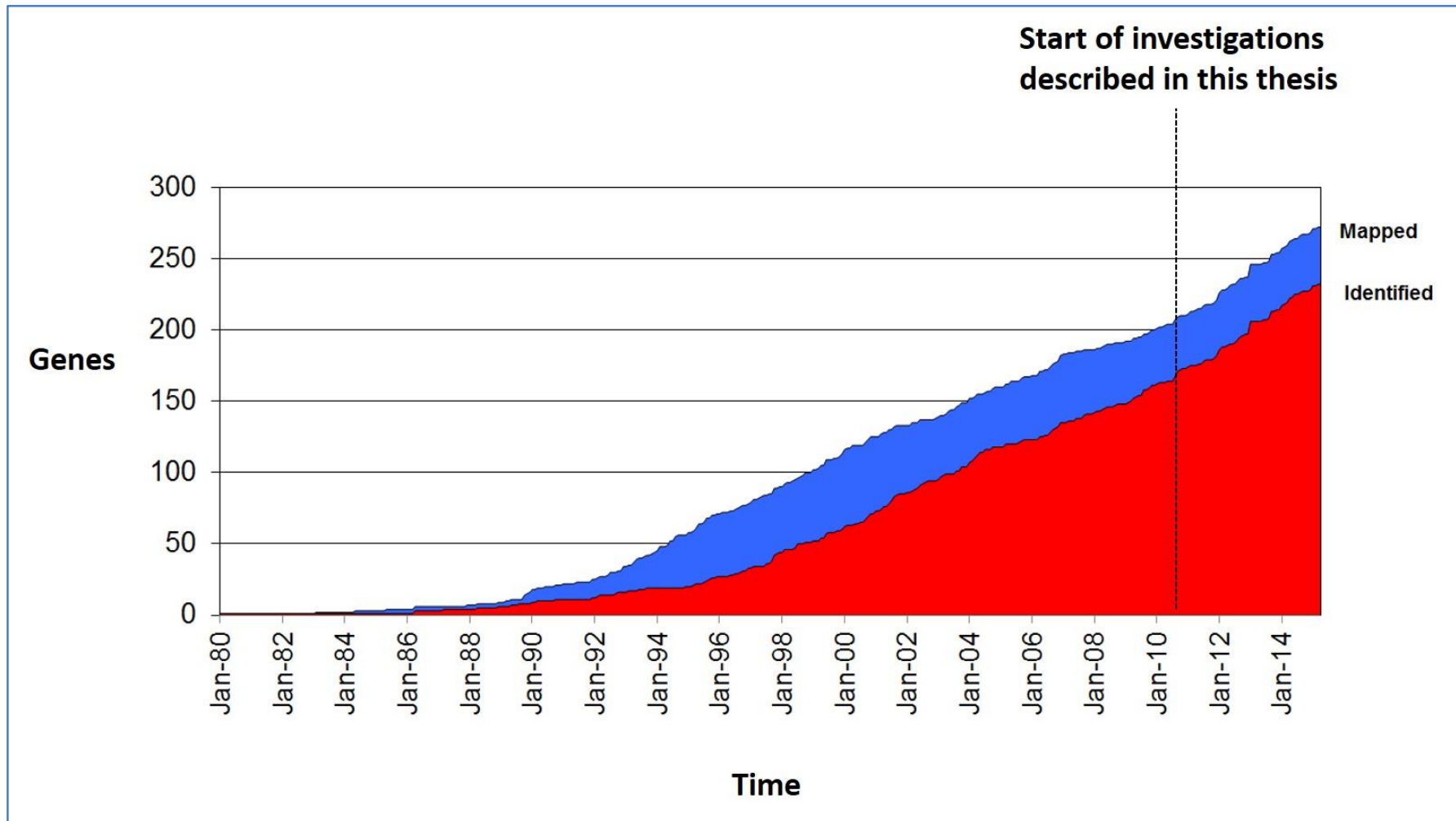


Figure 1.27 The number of mapped and identified retinal disease genes between 1980 and 2014. At the point where the investigations described in this began over 150 genes had been identified and over 200 had been mapped (Daiger *et al.* 1995). Figure adapted from RetNet: <https://sph.uth.edu/retnet/>

1.9 Clinical utility of next-generation sequencing with sequence capture

The decrease in sequencing costs brought about by NGS technologies was immediately recognised as a way to increase the molecular diagnostic yield in clinical applications for genetically heterogeneous disorders, and to reduce the overall cost and time to diagnosis (Rothberg and Leamon 2008).

Many different human diseases are potentially amenable to analysis using NGS, including cancers, single-gene disorders and multigenic disorders. NGS has also been shown to be highly effective in the rapid diagnosis of several human infections (Rothberg and Leamon 2008). However, at the start of the work presented here only a few human diseases had been investigated using NGS, and molecular diagnostics relied on sequential gene tests using Sanger sequencing. In the field of retinal genetics at the start of this project only a handful of genes were available for routine clinical testing including *RHO*, *PRPF31*, *RPGR*. The vast majority of genes were not routinely available for testing, even using Sanger sequencing. APEX arrays were offered as a testing service abroad just at the point that NGS became available on an experimental basis.

The work described in this thesis focussed on large cohorts of patients with genetic diseases without a molecular diagnosis, collected by clinicians in Oxford. This project was developed as a collaboration between the clinical and research teams to investigate NGS for use in clinical diagnostics. This included two projects, one on inherited retinal dystrophies (IRD), and the other on early onset epilepsies, both of which are described in this thesis.

Many genetic diseases including IRD and epilepsies show genetic heterogeneity; meaning that similar clinical phenotypes can be caused by mutations in many different genes. For example, nyctalopia (night blindness) and retinitis pigmentosa (“bone spicule pigmentation”) are clinical descriptions but do not provide a specific molecular diagnosis. Early onset epilepsies also show a high level of heterogeneity, and are discussed in the introduction to Chapter 5 p. 166.

Sequencing all known genes provides the possibility of making a molecular diagnosis in many more patients than is currently possible, which can provide information that is useful for a number of purposes, as discussed in the previous section.

Since sequencing such a large number of genes using conventional Sanger sequencing has been prohibitively costly and time consuming, NGS using sequence capture and clonal amplification offers the opportunity for cheaper, faster sequencing and a molecular diagnosis in a much larger number of patients.

Work by Morgan and colleagues (2010) evaluated the potential for Illumina sequencing in the diagnosis of Familial Breast Cancer and found that:

“clonal sequencing outperforms current diagnostic methods, providing a reduction in analysis times and in reagent costs compared with conventional sequencing” (Morgan et al. 2010).

For their study only a few regions of interest needed to be sequenced which could be covered by just two long PCR products, so the use of other sequence capture methods was not necessary.

1.9.1 Aims

The NGS technologies which emerged shortly before the start of this project offered the exciting possibility of expanding clinical genetic testing for retinal disorders in a way which had previously been unimaginable. However, before being able to introduce such technologies into clinical practice, the technologies needed to be evaluated in detail. The main aims of the work described in this project were:

- To demonstrate and evaluate the clinical utility of using Next-Generation Sequencing (NGS) and sequence capture methods for molecular diagnosis of IRDs in a clinical setting.
- To develop data filtering and analysis pipelines that can identify pathogenic variants from the vast amounts of raw data generated by NGS.
- To identify pathogenic variants in a cohort of IRD patients who have previously received no genetic testing.
- To obtain experimental data providing supporting or refuting evidence that mutations identified through NGS that are suspected to be pathogenic are the cause of the disease in two patients with early onset epilepsy.
- To find a molecular diagnosis for a cohort of patients with macular dystrophies, including patients with the specific phenotype of Stargardt disease, a recessive disease in which only one pathogenic variant has so far been discovered in some patients.

CHAPTER 2

Materials and Methods

2.1 Materials

2.1.1 Enzymes

The REDTaq enzyme and buffer used for polymerase chain reactions (PCR) was supplied by Sigma-Aldrich Ltd (UK). The deoxynucleotide triphosphates (dNTPs) used in PCR were supplied by BioLine (UK). Long Range PCR enzymes, buffer, dimethyl sulfoxide (DMSO), and dNTPs were supplied in the Expand Long Range dNTPack by Roche (UK). Exonuclease I (ExoI) and Shrimp Alkaline Phosphatase (SAP) were supplied by New England Biolabs (UK).

2.1.2 Solutions

TE Buffer (10mM Tris-Cl, 1mM EDTA, pH 8.0) was supplied by Promega (UK). TAE buffer (40mM Tris acetate, 1mM EDTA, pH 8.2) was supplied by Promega (UK). Molecular Biology grade water was supplied by Fisher Scientific Ltd (UK) and was purified by distillation, and treated to be DNase and RNase free. 100% ethanol was supplied by VWR chemicals (UK).

The QIAquick PCR Purification Kit contained Buffer PB, Buffer PE, Buffer EB, QIAquick spin columns, and collection tubes, all supplied by Qiagen (UK).

Agencourt AMPure XP beads used for Solid Phase Reversible Immobilisation (SPRI) were supplied by Beckman Coulter (UK).

2.1.3 Other materials

Hi-Res standard agarose was supplied by GeneFlow (UK). Ethidium bromide was supplied by Sigma-Aldrich Ltd (UK). O'Gene Ruler DNA Ladder Mix was supplied by Fisher Scientific Ltd (UK). 6x loading dye was supplied by Thermo Scientific Ltd (UK). Pipette tips in 10µl, 20µl, 200µl, and 1000µl sizes were supplied by StarLabs (UK). Microcentrifuge tubes were supplied by StarLab (UK). Bijoux tubes and 10ml tubes were supplied by Greiner Bio-one (UK). 96-well plates for PCR & Sanger sequencing were supplied by 4titude (UK). Nine litre container of compressed nitrogen gas was supplied by Praxair Surface Technologies Ltd (UK). DynaMag 2 magnet supplied by Life Technologies (UK).

2.1.4 Equipment and apparatus

2.1.4.1 Thermocycler

The G-Storm thermocycler was supplied by Bio-Rad (UK).

2.1.4.2 Electrophoresis

Electrophoresis tanks and power supplies were supplied by GeneFlow (UK) and Bio-Rad (UK).

Gel imaging equipment was supplied by Syngene (UK).

2.1.4.3 Bioanalyzer

The Agilent 2100 Bioanalyzer, chip priming station, DNA 7500 kits, and RNA 6000 Nano kits were all supplied by Agilent (UK).

2.1.4.4 Sanger sequencing

Sanger reactions were analysed by staff at the Department of Zoology at the University of Oxford using a 3730 DNA Analyzer with BigDye v3.1 chemistry, supplied by Applied Biosystems (UK).

2.1.4.5 Next-Generation Sequencing

Next-generation sequencing was performed using the 454 GS FLX sequencing machine with Titanium reagents, all supplied by Roche Diagnostics (UK).

An S220 Focussed-Ultrasonicator and Covaris microTUBE AFA Fiber snap-caps were supplied by Covaris (UK).

Illumina sequencing was performed on the HiSeq 2000 using library preparation reagents supplied by Illumina (UK) by the high-throughput sequencing group at the Wellcome Trust Centre for Human Genetics at the University of Oxford.

2.1.4.6 Sequence capture technologies

The NimbleGen sequence capture arrays, and hybridisation panel was provided by NimbleGen (UK), and capture performed by Morag Shanks at the Wellcome Trust Centre for Human Genetics at the University of Oxford.

SureSelect reagents were supplied by Agilent (UK).

DNA samples were sent to RainDance Technology (UK) where capture was performed using the RainDance Thunderstorm robot and RainDance reagents.

2.1.4.7 Spectrophotometry equipment

Qubit buffer, Qubit reagent, Qubit tubes, and the Qubit 2.0 fluorometer were all supplied by Life Technologies (UK).

The Nanodrop 1000 spectrophotometer was supplied by Thermo Scientific (UK).

2.1.1.8 SequalPrep Normalization Plates

A SequalPrep Normalization Plate (96) Kit containing SequalPrep normalization plates, SequalPrep Normalization Binding Buffer, SequalPrep Normalization Wash Buffer, and SequalPrep Normalization Elution Buffer comprised of 10mM Tris-HCl, pH 8.5, were all provided by Invitrogen (UK).

2.1.5 Software

A full list of software used in the work described in this thesis is provided in **Table 2.1**.

Table 2.1 List of software used.

Software name	Publisher	URL	Reference
Alamut Visual	Interactive Biosoftware	http://www.interactive-biosoftware.com/alamut-visual/	-
Alamut Batch	Interactive Biosoftware	http://www.interactive-biosoftware.com/alamut-batch/	-
BLAT	UCSC	https://genome.ucsc.edu/	Kent <i>et al.</i> 2002
Chromas	Technelysium	http://technelysium.com.au/?page_id=13	-
GS Reference Mapper	454	http://www.454.com/products/analysis-software/	-
GS Run Browser	454	-	-
In Silico PCR	UCSC	https://genome.ucsc.edu/	-
Integrative Genomics Viewer	Broad Institute	http://www.broadinstitute.org/igv/	Robinson <i>et al.</i> 2011
Ingenuity Variant Analysis	Ingenuity (Qiagen)	http://www.ingenuity.com/products/variant-analysis	-
Platypus	Wellcome Trust Centre for Human Genetics, University of Oxford	http://www.well.ox.ac.uk/platypus	Rimmer <i>et al.</i> 2014
PolyPhen2	Sunyaev lab, Harvard	http://genetics.bwh.harvard.edu/pph2/	Adzhubei <i>et al.</i> 2010
Primer3	Whitehead Institute for Biomedical Research	http://primer3.sourceforge.net/	Untergasser <i>et al.</i> 2012 and Koressaar <i>et al.</i> 2007
Stampy	Wellcome Trust Centre for Human Genetics, University of Oxford	http://www.well.ox.ac.uk/stampy	Lunter and Goodson 2011
SureDesign	Agilent (UK)	https://earray.chem.agilent.com/suredesign/	-
UCSC Genome Browser	UCSC	https://genome.ucsc.edu/	Kent <i>et al.</i> 2002

2.2 Methods

2.2.1 Patient cohorts and ophthalmological examination

Patients for this project were recruited by Susan Downes in the Oxford Eye Hospital. Consent for testing was obtained according to the declaration of Helsinki, and ethics was approved by Essex REC, reference number 08/H0302/96.

Patients recruited had a variety of clinical phenotypes. These included:

- Maculopathy
- Cone-Rod Dystrophy
- Retinitis pigmentosa

There were a total of 100 patients in the cohort, of whom 57 had a maculopathy. Of these 57 patients, 24 were used for the comparison analysis of technologies (see Chapter 3), and 12 were used as positive controls in the NGS screen of patients with an unknown molecular diagnosis (see Chapter 4). The other 43 patients in the cohort had either a cone-rod dystrophy or retinitis pigmentosa. These patients had not had any prior genetic testing and were the “unknowns” for which a molecular diagnosis was sought in Chapter 4.

2.2.2 DNA extraction and dilution

DNA extraction was performed by the Oxford Regional Genetics Laboratory (ORGL) using Qiagen Autopure LS and EZ1 systems supplied by Qiagen (UK). DNA was extracted from whole blood samples. DNA samples underwent multiplex ligation probe amplification (MLPA) at the ORGL using kits supplied by MRC Holland to test for copy number variants and exon deletion. MLPA involves the ligation of probes to a target sequence of DNA followed by amplification of the probes using PCR. Only when both of a pair of probes ligate to a sequence can the amplification proceed, which allows the detection of copy number variants and other changes to sequence (Schouten *et al.* 2002).

Samples provided by the ORGL were diluted to a concentration of approximately 200ng/μl in TE buffer. Tubes of DNA received from the ORGL were rotated on a Labquake benchtop rotator for 90 minutes, then quantified using Qubit fluorometry (described in section 2.2.3.2 p. 77) and stored at 4°C until needed.

2.2.3 Nucleic acid quantitation

Spectrophotometry is a method of measuring the composition of a substance by detecting the level of absorbance of different wavelengths of light that pass through it. Spectrophotometry can be used to determine the concentration of nucleic acids present in a sample by measuring the absorbance of UV light at 260nm using a photo-detector.

When stimulated by light at a 260nm wavelength pure double-stranded DNA has an attenuation coefficient of 0.020 (μg/ml)/cm, single stranded DNA 0.027 (μg/ml)/cm, and single-stranded RNA 0.025 (μg/ml)/cm. The optical density or absorbance value of a nucleic acid can be determined by first measuring a blank sample containing the same buffer as the samples. The samples are then measured, and the two readings used in the following equations to determine what has been absorbed by the samples alone:

$$T = \frac{I}{I_0}$$

Where T= Transmittance, I = the intensity of the light passing through the sample, and I₀ = the intensity of the light passing through the blank sample measurement.

$$A = -\log_{10} T$$

Where A = the absorbance measurement, and T = the transmittance value calculated above.

Higher concentrations of nucleic acid absorb more light, so by determining the absorbance value and comparing it to recordings made at known concentrations, it is possible to accurately determine the concentration of a nucleic acid in a solution.

Spectrophotometry can also be used to determine the purity of a nucleic acid sample, by comparing the absorbance at two different wavelengths; 260nm and 280nm. Protein has a maximum absorbance at 280nm and nucleic acids at 260nm. The 260/280 ratio of absorbance is expected to be 2.0 for pure DNA, and 1.8 for pure RNA. 260/230 ratios can also be used to determine if substances other than protein have contaminated a nucleic acid sample. DNA samples were quantified using a Qubit 2.0 fluorometer (2.2.3.1) or a nanodrop (2.2.3.2).

2.2.3.1 Spectrophotometry using the Nanodrop apparatus

A nanodrop 1000 spectrophotometer (Thermo Scientific) was used to determine the concentration of RNA samples, and also to determine the purity of both DNA and RNA samples by measurement of 260/280 and 260/230 ratios. Manufacturer's instructions were followed with 1µl of sample used for each reading.

2.2.3.2 Spectrophotometry using the Qubit fluorometric quantitation method

DNA sample quantitations were performed using a Qubit 2.0 fluorometer and dsDNA BR assay kit reagents supplied by Life technologies (UK). A Qubit working solution was prepared by making a 1:200 dilution of the Qubit reagent (a dye that binds to double stranded DNA with fluorescence that is excited at 510nm and has an emission of 527nm) into Qubit buffer. 200µl of working solution was prepared for each of the samples, plus the two standards. Qubit tubes were prepared containing 190µl of working solution and 10µl of standard for each of the two standard tubes. For each of the sample tubes 199µl of Qubit working solution and 1µl of sample were prepared in a Qubit tube. Each of the tubes was vortexed, then stored for two minutes under a foil cover to prevent degradation of the dye by light. The Qubit 2.0 fluorometer was calibrated using the two standard tubes. The absorbance values measured for samples by the machine were multiplied by 200 to give the concentration of the original undiluted DNA sample.

2.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis allows DNA fragments, such as those generated by PCR to be separated by size. The basic principle is that shorter fragments of negatively charged DNA will be carried further through an agarose gel by an electric current over the same amount of time compared to longer fragments which encounter more resistance. Fragments of unknown size can then be compared to a ladder of fragments of known size to determine the length of the DNA sequence in base pairs. The concentration of agarose in the gel can be changed, with different percentages used to optimally resolve different sized products.

1-2% agarose gels were made using Hi-Res Standard Agarose from GeneFlow (UK) dissolved in 1x TAE using a standard microwave oven. When cooled ethidium bromide solution was added to a concentration of 0.5µg/ml, after which the agarose was poured into a Bio-Rad tray mould with a comb and left to set.

Agarose gel electrophoresis was performed using 1x TAE buffer. 5µl of DNA marker ladder at 0.5µg/µL concentration was also used. For most purposes 3µl of PCR product would be mixed with 5µl of loading dye and loaded into each sample well.

Gels were usually run at 100V for one hour and viewed on a U:Genius Syngene trans-illuminator provided by Syngene (UK), which uses UV light to stimulate the ethidium bromide-stained DNA, making it visible. Photographs of the gel were taken using the 3-megapixel camera mounted on the transilluminator and saved to a UDB disc.

2.2.5 Agilent 2100 Bioanalyzer

The Agilent 2100 Bioanalyzer allows for sizing and quantitation of nucleic acids and proteins on a chip, and relies on the same principle as agarose gel electrophoresis. The chip is made up of wells for samples, gel, and ladders, all containing a sieving polymer and fluorescent dye. Micro-channels connect these wells. A set of 16 pin electrodes controlled by independent power

supplies fit into the wells when the device is closed. Protein and nucleic acid separation is electrophoretically driven by a voltage gradient through the pin electrodes. The constant mass to charge ratio applied to the biomolecules as they pass through the sieving polymer allows them to be separated by size. The data can then be digitally produced as an electropherogram or as a densitometry plot.

The Agilent 2100 Bioanalyzer was used for sizing DNA using DNA 7500 chips and reagents, and RNA using RNA Pico 6000 chips and reagents. In each case the manufacturer's protocols were followed without deviation. Protocols are available at <http://www.genomics.agilent.com>.

2.2.6 Polymerase Chain Reaction (PCR)

The Polymerase Chain Reaction (PCR) is a technique used commonly in molecular biology to amplify a specific sequence of DNA. The process begins by denaturing a dsDNA sample that is used as a template by heating it to approximately 95°C. The temperature is then lowered to approximately 58°C, at which point oligonucleotide primers (short single stranded sequences of DNA designed to be complementary to the target sequence) then bind to the single stranded template DNA (known as annealing). The temperature is then raised again to approximately 72°C to allow extension of the double strand to occur. DNA replication then proceeds from the most 3' prime base of double-stranded sequence, with a heat-resistant DNA polymerase enzyme isolated from *Thermophilus aquaticus* (*Taq* enzyme) incorporating a single nucleotide at a time along the strand, until the target region for amplification has once again become double stranded. This creates two double stranded DNA molecules from the initial template. The reaction is then usually repeated a number of times, with the amount of target DNA doubling each time, as each DNA amplicon created can then act as a template in the next cycle of denaturation, annealing, and extension. Amplicons generated through PCR can then be used for a number of applications, including Sanger sequencing. The success of PCR amplification can be determined by agarose gel electrophoresis.

2.2.6.1 Primer design

Primers for PCR and Sanger sequencing were designed using Primer3 (Koressaar and Remm 2007, Untergasser *et al.* 2012), available at <http://primer3.sourceforge.net/>. *In silico* PCR programs based on BLAT (Kent 2002), available at <http://genome.csdb.cn/cgi-bin/hgPcr>, were used to predict whether primers would amplify the correct regions. Primers were synthesised by Sigma-Aldrich (UK) and diluted to a concentration of 10µM for PCR and 3.5µM for Sanger sequencing.

Primers were designed to fulfil the following criteria:

- Similar GC content for primer pairs.
- Minimal secondary structure to reduce self-complementarity.
- Low complementarity of sequence between primers to prevent primer-dimer formation.
- Random base distribution.
- Anchoring pair of C or G nucleotides at one end of each primer.
- Oligonucleotide length of at least 20 nucleotides, for greater specificity.
- Less than 4°C annealing temperature difference between primer pairs.
- Where possible primers were selected with an annealing temperature close to 58°C to allow multiple amplifications to be performed on the same thermocycler run.

Table 2.2 Validation primers.

Primer name	Gene targeted	Primer sequence
C5orf42-F	<i>C5orf42</i>	AAAGGAAGTGAATGCCAGTGA
C5orf42-R	<i>C5orf42</i>	TGCATTCTATCCAGAAAGTGAGG
PIGQ-F	<i>PIGQ</i>	CCAGGCTCCGTGTTAGTCAC
PIGQ-R	<i>PIGQ</i>	CAGCCCTGCAGAGACCAG
PIGQ-2F	<i>PIGQ</i>	CACGCAGTGAGGTGCTCTT
PIGQ-5R	<i>PIGQ</i>	GGGGACATGAGGTGGATGTA
CBL-8F	<i>CBL</i>	GAGATGGGCTCCACATTCC
CBL-1R	<i>CBL</i>	GAACTTGGGGCAGATACTGG
D942N-F	<i>D942N</i>	GGAATCAGGGTTTCTGGACA
D942N-R	<i>D942N</i>	GGAGGACCTGAAGAAGGACA
M1808T-F	<i>M1808T</i>	GGGAGCTCTGGGGACTATG
M1808T-R	<i>M1808T</i>	GGATGACGCCCTGAGAA
CRX R41W F	<i>CRX</i>	GTGAGGTGTAGAAGGGCAGG
CRX R41W R	<i>CRX</i>	CATCACTCTTTGTTCCGGGC
ABCA4_splice01F	<i>ABCA4</i>	GGGGAAAATGTTCTGAGTTGC
ABCA4_splice01R	<i>ABCA4</i>	CGTGGACCCCTGAGACTG
ABCA4_splice02F	<i>ABCA4</i>	AATGTCCCTGCAAAGATGG
ABCA4_splice02R	<i>ABCA4</i>	GTTTGCAGCCACCCAAAC
ABCA4_splice03F	<i>ABCA4</i>	GCAGGATGCAGACCTAGGAG
ABCA4_splice03R	<i>ABCA4</i>	ATGAGTTTCAGGCCCCAGTT

2.2.6.2 Standard PCR

The following master mix was used for PCR amplification of fragments up to 1000bp in size:

DNA (50ng/μl)	0.5μl
Primers (pair at 10μM)	0.2μl
dNTPs (10mM)	0.5μl
10x REDTaq Buffer (100mM Tris-HCl pH 8.3, 500mM KCl, 11mM MgCl ₂ , and 0.1% gelatin)	2.5μl
REDTaq (1 unit/μl in 20mM Tris-HCl pH 8.0)	0.2μl
H ₂ O (molecular biology grade)	21.1μl
Total	25μl

The following thermocycler program was used:

95°C	4 minutes
30x cycles of:	
95°C	30 seconds
58°C	30 seconds
72°C	1 minute
End cycle	
72°C	7 minutes

The primers in **Table 2.2** were used for the standard PCR protocol and for Sanger sequencing (described in section 2.2.8.1 p. 87) to validate variants identified by NGS in results chapters 4, 5, and 6.

2.2.6.3 Long Range PCR with Illumina HiSeq 2500 sequencing

Primers were designed for Long Range sequencing of the *ABCA4* gene. Initial experiments showed that 10kbp was the maximum size of amplification that could successfully be produced consistently. The gene was therefore divided into sections of approximately 10kbp or less with primers designed to bind to exons where possible, and for amplifications to overlap where possible as well, to generate contiguous amplification of the entire gene. The primers used are listed in **Table 2.3** and the location of the products they generated relative to the gene are shown in **Figure 6.2**. Primers were also designed to amplify the 5' and 3' UTRs and four regions which ENCODE data suggested could contain potential promoter regions, based on the presence of H3K27AC marks (Creyghton *et al.* 2010).

The Long Range amplifications were performed using Roche Long Range Expand dNTPack using the following mastermix:

DNA (20ng/μl)	2.0μl
Primers (pair at 10μM)	1.5μl
dNTP mix (each at 10mM)	1.25μl
DMSO (100%)	0.75μl
5x Buffer (with 12.5mM MgCl ₂)	5μl
Enzyme mix (1 Unit/μl)	0.35μl
H ₂ O (molecular biology grade purity)	14.15μl
Total	25μl

The following thermocycler program was used:

94°C 2 minutes

35x cycles of:

94°C 10 seconds

58°C 30 seconds

68°C 17 minutes

End cycle

68°C 7 minutes

Table 2.3 ABCA4 Long Range-PCR primers.

Primer name	Primer sequence (5' -> 3')	Primer name	Primer sequence (5' -> 3')
ABCA4 Exon 1 to 3 F	GTGGGGCTTTGAAAACAGG	ABCA4 Exon 1 to 3 R	TCGTGTGGTCATTAGCATGG
ABCA4 Exon 2 to 5 F	GAGAGTTGATCAGAAGGTAGACC	ABCA4 Exon 2 to 5 R	TCTTGATCTGGTTAAGGAATGCC
ABCA 4 Exon 4 to 6 F	GAAGAGCTTGAAGAAGTCCACG	ABCA4 Exon 4 to 6 R	AAGAACTCCTCATGAATGCACC
ABCA4 Exon 6 to 6.5 F	GGGCTCTCCTCTTCTTTTGG	ABCA4 Exon 6 to 6.5 R	AGCGCTTCATCATCTTCAGC
ABCA4 Exon 6.5 to 7 F	ACGGCTGTCTAGGAGTGTGG	ABCA4 Exon 6.5 to 7 R	CCCTTCACAGCCTTATCTTCC
ABCA4 Exon 7 to 11 F	AATACCTTCTTCACCAAGCTGC	ABCA4 Exon 7 to 11 R	TCAATCTGAGATCTTG GGGAGG
ABCA4 Exon 11 to 11.5 F	CCCATCCATTTCTCTTGAGC	ABCA4 Exon 11 to 11.5 R	AGGCAGCTTGGTGAAGAAGG
ABCA4 Exon 11.5 to 12 F	GTAGAGAGAGGGCACGTTGG	ABCA4 Exon 11.5 to 12 R	AAAGGCCTTCCCAGTCTAGC
ABCA4 Exon 12 to 16 F	ACTCAGTGCCAAATCCAAATGC	ABCA4 Exon 12 to 16 R	CTGGAGGAAAACATGTTCTGGG
ABCA4 Exon 16 to 19 F	GTAGAAGGTGATGTTTACAGACGG	ABCA4 Exon 16 to 19 R	TTTGGATTTGGCACTGAGTACC
ABCA4 Exon 19 to 24 F	CTTGTTCTGGAGTTAGGTCATCG	ABCA4 Exon 19 to 24 R	CGTCTGAACATCACCTTCTACG
ABCA4 Exon 23 to 27 F	GGTGTGTTGGAATCTCTTGACC	ABCA4 Exon 23 to 27 R	GGAAGGCTCTACTGCTCAGG
ABCA4 Exon 27 to 31 F	TCTTGTAGAATTTCCGTGCTGC	ABCA4 Exon 27 to 31 R	GGTCAAGAGATTCCAACACACC
ABCA4 Exon 31 to 36 F	GGTGCTTGGATTTGTTTACC	ABCA4 Exon 31 to 36 R	GGAACATCTCCGACTTCTTGG
ABCA4 Exon 36 to 38 F	CTCAGTGCAAGGTCAATGAGG	ABCA4 Exon 36 to 38 R	TGAATTATTCCGTGAGTGCTGG
ABCA4 Exon 38 to 43 F	CTTGAATGTGGTTGTTTTGCCG	ABCA4 Exon 38 to 43 R	CATCCTTCCTGTTTGATGTCCC
ABCA4 Exon 43 to 47 F	GAATGGTGCCCATACATCG	ABCA4 Exon 43 to 47 R	CACATTCAAGATGCTCACTGG
ABCA4 B3A E46-50 F	CGGGGTTTCTAGTTCTGGGG	ABCA4 B3A E46-50 R	CATCATCAGAGAAGGGAGGGC
Promoter Region 1 F	ACCAGCAAGTTCCAGTCAGG	Promoter Region 1 R	CCAAGACCACTTTTGAAAAACC
Promoter Region 2 F	GGTTTTTCAAAAAGTGGTCTTGG	Promoter Region 2 R	CAGCCAGAAAGGAACTCTGG
Promoter Region 3 F	GCTAATGACCACACGAAGACC	Promoter Region 3 R	AAAATTGAGCTATAGGCATTTTAGG
Promoter Region 4 F	CCCTCCAAATACCTTCACAGC	Promoter Region 4 R	GCCAGAAGTCTTCCTGATGG
5' UTR F	TGATCTCTCCAGTG GCTTGC	5' UTR R	TCATAGCCTTCACTCCAGCC
3' UTR F	AACCAAGGACCAACTCTCGC	3' UTR R	GTCAGGCCTTTGTCCAGTCC

All primers were designed to have an annealing temperature of 58°C.

2.2.7 DNA purification

2.2.7.1 QIAquick PCR Purification Kit

PCR products were cleaned up using the QIAquick purification kit (Qiagen, UK) following manufacturer's protocols but briefly, five volumes of binding buffer were added for each volume of sample. These were mixed, pipetted into a QIAquick spin column inside a collection tube, and centrifuged at $17,900 \times g$ for one minute. The flow-through was discarded and 750 μ l of wash buffer added, and the column centrifuged at $17,900 \times g$ for one minute. The flow-through was discarded and this wash step repeated. The spin column was transferred to a fresh collection tube and 50 μ l of elution buffer pipetted onto the centre of the column. The column was incubated for one minute before being centrifuged at $17,900 \times g$ for one minute.

The principle of these purification methods is that high salt concentrations in the buffers cause binding of DNA to a silica membrane in the spin column. The column is centrifuged, which forces the solution through the silica membrane in the spin column, allowing the DNA to bind while other contaminants pass through. Subsequent wash steps using buffer PE further remove any contaminant substances. The purified DNA is then eluted in a salt free elution buffer of Tris-Cl or molecular biology grade H₂O.

The exact composition of buffers PB and PE are confidential but they contain a high concentration of guanidine hydrochloride and isopropanol. Buffer PE requires the addition of 100% ethanol before use. Buffer EB is composed of 10 mM Tris-Cl, pH 8.5.

2.2.7.2 Solid Phase Reversible Immobilisation (SPRI) beads

Solid Phase Reversible Immobilisation (SPRI) beads were used to purify DNA samples, and to size select DNA libraries by removing fragments of less than 300bp in size. SPRI beads are microscopic polystyrene beads, surrounded by magnetite, coated in a layer of carboxyl molecules. The process requires an equal volume of SPRI beads in Poly-Ethylene Glycate (PEG) solution to be added to a volume of DNA to be purified, then vortexed. The DNA binds to the surface of the beads in the presence of the PEG solution. The tubes were placed in a

magnetic field that caused the beads to attach to the side of the microcentrifuge tube, due to the paramagnetic properties of the magnetite. The solution containing PEG and other contaminants was then washed away by a buffer containing ethanol. Beads were then resuspended in TE buffer and vortexed. Beads were again pulled to the side of the tube by a magnet, and the TE buffer containing DNA was transferred to a fresh tube.

2.2.8 Sequencing

2.2.8.1 Sanger sequencing

PCR products to be sequenced were first cleaned up using an ExoSAP reaction where an exonuclease (Exo) enzyme is used to remove primers from the PCR, while the Shrimp Alkaline Phosphatase (SAP) removes the dNTPs.

2.2.8.1.1 ExoSAP

The following recipe was used to create a mastermix:

ExoI (20 Units/ μ l)	0.1 μ l
SAP (1 Unit/ μ l)	1 μ l
10x SAP buffer	1 μ l
ddH ₂ O	0.9 μ l
Total	3μl

3 μ l of mastermix was added to each PCR product. The following program was used on a thermocycler:

37°C for 40 minutes
80°C for 15 minutes

2.2.8.1.2 Sanger reaction

The following mastermix was used to create a Sanger sequencing reaction, one each for the forward and reverse primer for each PCR product:

DNA (PCR product)	1.0µl
Primer (3.2µM)	0.5µl
2.5x BigDye	1.0µl
5x BigDye buffer	1.5µl
H ₂ O	6.0µl
Total	10µl

The following thermocycler program was used:

96°C	1 minute
35x cycles of:	
96°C	30 seconds
50°C	15 seconds
60°C	4 minutes
End cycle	

2.2.8.1.3 Ethanol precipitation

Sanger reactions were cleaned up using an ethanol precipitation. The following reaction mixture was prepared for each sample:

100% ethanol	50µl
Sodium acetate (3M)	2µl
125mM EDTA	2µl
Total	54µl

54µl of the reaction mixture was added to each Sanger reaction and the mixture was incubated at room temperature for 15 minutes. The plate was then centrifuged at 4°C for 30 mins at 3000 x *g*. The plate was turned upside down onto a paper towel, then spun briefly upside down onto a paper towel to drain any remaining solution for ten seconds at 180 x *g*. 70µl of 70% ethanol was added to each sample, and the plate was spun at 4°C for 15 mins at 1650 x *g*. The plate was again turned upside down onto a paper towel and spun briefly for ten seconds at 180 x *g* to drain any remaining solution. The plate was allowed to air dry at room temperature for approximately one hour before being analysed.

Sanger reactions were run out on an Applied Biosystems 3730xl DNA Analyzer by the Department of Zoology at the University of Oxford to generate electropherograms. The data were analysed using Chromas Lite software (**Table 2.1**) supplied by Technelysium (Australia) to identify single nucleotide polymorphisms and other genetic variations.

2.2.8.2 Sequence capture and enrichment

A targeted sequence capture design was created by Dr Morag Shanks for all of the sequence capture and enrichment technologies used in this project. The same design was used for each technology, which targeted all the known coding exons of 74 genes known to cause retinal disease when mutated. This is described in full in Shanks *et al.* (2012). But briefly, these genes were selected by searching the literature for publications that showed strong evidence the gene was involved in a known retinal process.

Only 74 genes were chosen for two reasons; firstly, the initial capture design was made at a time when less than 100 genes had been identified, so it was not possible to target significantly more than this. In addition, there were technological limitations imposed on the design of NimbleGen sequence capture arrays that prevented more genes being targeted.

In total there were 1355 regions of interest used in the capture design, covering a total of 463,349 bases across 74 genes. The full list of genes targeted, the proteins they encode, and the diseases they are known to be associated with are shown in **Table 2.4**.

Table 2.4 Genes used in targeted capture.

Gene	Encodes	Associated with
<i>ABCA4</i>	ATP-binding cassette, sub-family A (ABC1), member 4	Stargardt disease (OMIM 248200), Retinitis pigmentosa 19 (OMIM 601718), Cone-rod dystrophy 3 (OMIM 604116), Age-related macular degeneration 2 (OMIM 153800)
<i>AHI1</i>	Abelson helper integration site 1	Joubert syndrome-3 (OMIM 608629)
<i>AIPL1</i>	Aryl hydrocarbon receptor-interacting protein-like 1	Retinitis pigmentosa, juvenile, AIPL1-related (OMIM 604393)
<i>BEST1</i>	Bestrophin 1	Vitelliform macular dystrophy 2 (OMIM 153700), and other retinopathies
<i>C1QTNF5</i>	C1q and tumour necrosis factor- related 5 (CTRP5)	Autosomal dominant late-onset retinitis pigmentosa (OMIM 605670)
<i>CA4</i>	Carbonic anhydrase IV	Retinitis pigmentosa 17 (OMIM 600852)
<i>CACNA2D4</i>	Calcium channel, voltage-dependent, alpha 2/delta subunit 4	Retinal cone dystrophy 4 (OMIM 610478)
<i>CC2D2A</i>	Coiled-coil and C2 domains-containing protein 2A	COACH syndrome (OMIM 216360), Joubert syndrome 9 (OMIM 612285), Meckel syndrome 6 (OMIM 612284)
<i>CDH23</i>	Cadherin-related 23	Usher syndrome, type 1D (OMIM 605516)
<i>CEP290</i>	Centrosomal protein 290-KD	Bardet-Biedl syndrome 14 (OMIM 615991), Joubert syndrome 5 (610188), Leber congenital amaurosis 10 (OMIM 611755), Meckel syndrome 4 (OMIM 611134), Senior-Loken syndrome 6 (OMIM 610189)
<i>CERKL</i>	Ceramide kinase-like	Retinitis pigmentosa 26 (OMIM 608380)
<i>CLRN1</i>	Clarin 1	Retinitis pigmentosa 61 (OMIM 614180)
<i>CNGA1</i>	Cyclic nucleotide gated channel alpha 1	Retinitis Pigmentosa 49 (OMIM 613756)
<i>CNGB1</i>	Cyclic nucleotide gated channel beta 1	Retinitis Pigmentosa 45 (OMIM 600724)
<i>CNGB3</i>	Cyclic nucleotide gated channel beta 3	Achromatopsia-3 (OMIM 262300), Juvenile macular degeneration (OMIM 248200)
<i>CRB1</i>	Crumbs <i>Drosophila</i> homolog 1	Leber congenital amaurosis 8 (OMIM 613835), Pigmented paravenous chorioretinal atrophy (OMIM 172870), Retinitis pigmentosa-12 (OMIM 600105)
<i>CRX</i>	Cone-rod homeobox	Cone-rod retinal dystrophy-2 (OMIM 120970), Leber congenital amaurosis 7 (OMIM 613829)
<i>DFNB31</i>	Deafness, autosomal recessive 31	Autosomal recessive non-syndromic deafness (OMIM 607928), Usher Syndrome
<i>EFEMP1</i>	EGF containing fibulin-like extracellular matrix protein 1	Doyne honeycomb degeneration of retina (OMIM 126600)
<i>ELOVL4</i>	Elongation of very long chain fatty acids-like 4	Stargardt disease 3 (OMIM 600110)
<i>EYS</i>	Eyes shut <i>Drosophila</i> homolog	Retinitis pigmentosa 25 (OMIM 602772)

Table 2.4 continued

Gene	Encodes	Associated with
<i>FSCN2</i>	Fascin Sea urchin homolog 2	Retinitis pigmentosa 30 (OMIM 607921)
<i>GPR98</i>	G protein-coupled receptor 98	Usher syndrome 2C (OMIM 605472)
<i>GUCA1A</i>	Guanylate cyclase activating protein 1	Cone dystrophy 3 (OMIM 602093)
<i>GUCY2D</i>	Guanylate cyclase 2D	Cone-rod dystrophy 6 (OMIM 601777), Leber congenital amaurosis 1 (OMIM 204000)
<i>IDH3B</i>	Isocitrate dehydrogenase 3	Retinitis pigmentosa 46 (OMIM 612572)
<i>IMPDH1</i>	Inosine 5'-monophosphate dehydrogenase-like 1	Leber congenital amaurosis 11 (OMIM 613837), Retinitis pigmentosa 10 (OMIM 180105)
<i>KCNV2</i>	Potassium channel, voltage gated modifier subfamily V, member 2	Retinal cone dystrophy 3B (OMIM 610356)
<i>LRAT</i>	Lecithin retinol acyltransferase	Leber congenital amaurosis 14 (OMIM 613341)
<i>MERTK</i>	MER proto-oncogene tyrosine kinase	Retinitis pigmentosa (OMIM 613862)
<i>MYO7A</i>	Myosin VIIA	Usher syndrome type 1B (OMIM 276900)
<i>NPHP1</i>	Nephronophthisis 1	Senior-Loken syndrome-1 (OMIM 266900)
<i>NR2E3</i>	Nuclear receptor subfamily 2, group E, member 3	Enhanced S-cone syndrome (OMIM 268100), Retinitis pigmentosa 37 (OMIM 611131)
<i>NRL</i>	Neural retina leucine zipper	Retinal pigmentosa 27 (OMIM 613750)
<i>PCDH15</i>	Protocadherin-related 15	Usher syndrome type 1D/F digenic & type 1F (OMIM 601067 & 602083)
<i>PDE6A</i>	Phosphodiesterase 6A, cGMP-specific, rod, alpha	Retinitis pigmentosa 43 (OMIM 613810)
<i>PDE6B</i>	Phosphodiesterase 6B, cGMP-specific, rod, beta	Congenital stationary night blindness 2 (OMIM 163500), Retinitis pigmentosa 40 (OMIM 613801)
<i>PITPNM3</i>	PYK2 N-terminal domain-interacting receptor 1	Cone-rod dystrophy 5 (OMIM 600977)
<i>PRCD</i>	Progressive rod-cone degeneration	Retinitis pigmentosa 36 (OMIM 610599)
<i>PROM1</i>	Prominin mouse homolog 1	Cone-rod dystrophy 12 (OMIM 612657), Macular dystrophy 2 (OMIM 608051), Retinitis pigmentosa 41 (OMIM 612095), Stargardt disease 4 (OMIM 603786)
<i>PRPF3</i>	Pre-mRNA processing factor <i>S. Cerevisiae</i> homolog 3	Retinitis pigmentosa 18 (OMIM 601414)
<i>PRPF31</i>	Pre-mRNA processing factor <i>S. Cerevisiae</i> homolog 31	Retinitis pigmentosa 11 (OMIM 600138)
<i>PRPF8</i>	Pre-mRNA processing factor <i>S. Cerevisiae</i> homolog 8	Retinitis pigmentosa 13 (OMIM 600059)
<i>PRPH2</i>	Peripherin mouse homolog 2	Choroidal dystrophy central areolar 2 (OMIM 613105), Leber congenital amaurosis 18 (OMIM 608133), Patterned macular dystrophy 1 (OMIM 616915), Vitelliform macular dystrophy 3 (OMIM 608161), Retinitis punctata albescens (OMIM 136880)
<i>RAX2</i>	Retina and anterior neural fold homeobox 2	Cone-rod dystrophy 11 (OMIM 610381), Age-related macular degeneration 6 (OMIM 613757)

Table 2.4 continued

Gene	Encodes	Associated with
<i>RBP3</i>	Retinol binding protein 3, interstitial	Retinitis pigmentosa 66 (OMIM 615233)
<i>RDH12</i>	Retinol dehydrogenase 12	Leber congenital amaurosis 13 (OMIM 612712)
<i>RDH5</i>	Retinol dehydrogenase 5	Fundus albipunctatus (OMIM 136880)
<i>RGR</i>	Retinal G protein coupled receptor	Retinitis pigmentosa 44 (OMIM 613769)
<i>RHO</i>	Rhodopsin	Congenital stationary night blindness 1 (OMIM 610445), Retinitis Pigmentosa 4 (OMIM 613731), Retinitis punctate albescens (OMIM 136880)
<i>RIMS1</i>	Regulating synaptic membrane exocytosis 1	Cone-rod dystrophy 7 (OMIM 603649)
<i>RLBP1</i>	Retinaldehyde-binding protein 1	Bothnia retinal dystrophy (OMIM 607475), Fundus albipunctatus (OMIM 136880), Newfoundland rod-cone dystrophy (OMIM 607476), Retinitis pigmentosa albescens (OMIM 136880)
<i>ROM1</i>	Retinal outer segment membrane protein 1	Retinitis pigmentosa 7 (OMIM 608133)
<i>RP1</i>	Retinitis pigmentosa 1 (autosomal dominant)	Retinitis pigmentosa 1 (OMIM 180100)
<i>RP2</i>	Retinitis pigmentosa 2 (X-linked recessive)	Retinitis pigmentosa 2 (OMIM 312600)
<i>RP9</i>	Retinitis pigmentosa 9 (autosomal dominant)	Retinitis pigmentosa 9 (OMIM 180104)
<i>RPE65</i>	Retinal pigment epithelium-specific protein 65kDa	Leber congenital amaurosis 2 (OMIM 204100)
<i>RPGR</i>	Retinitis pigmentosa GTPase regulator	X-linked cone-rod dystrophy 1 (OMIM 304020), X-linked atrophic macular degeneration (OMIM300834), Retinitis pigmentosa 3 (OMIM 300029), X-linked retinitis pigmentosa (OMIM 300455)
<i>RPGRIP1L</i>	RPGRIP1-like	Possible modifier of retinal degeneration in patients with ciliopathies due to other mutations
<i>SAG</i>	S-antigen; retinal and pineal gland (arrestin)	Oguchi disease (OMIM 258100), Retinitis pigmentosa 47 (OMIM 613758)
<i>SEMA4A</i>	Sema domain, immunoglobulin domain, transmembrane domain an short cytoplasmic domain, (semaphorin) 4A	Cone-rod dystrophy 10 (OMIM 610283), Retinitis pigmentosa 35 (OMIM 610282)
<i>TIMP3</i>	TIMP metalloproteinase inhibitor 3	Sorsby fundus dystrophy (OMIM 136900)
<i>TMEM67</i>	Transmembrane protein 67	Bardet-Biedl syndrome modifier (OMIM 209900)
<i>TOPORS</i>	Topoisomerase I binding, arginine/serine-rich, E3 ubiquitin protein ligase	Retinitis pigmentosa 31 (OMIM 609923)
<i>TTC8</i>	Tetratricopeptide repeat domain-containing protein 8	Retinitis pigmentosa 51 (OMIM 613464), Bardet-Biedl Syndrome 8 (OMIM 615985)
<i>TULP1</i>	Tubby like protein 1	Leber congenital amaurosis 15 (OMIM 613843), Retinitis pigmentosa 14 (OMIM 600132)

Table 2.4 continued

Gene	Encodes	Associated with
<i>UNC119</i>	Unc 119 <i>C. elegans</i> homolog	Cone-rod dystrophy (unconfirmed), expressed in photoreceptors in retina
<i>USH1C</i>	Usher syndrome 1C (autosomal recessive, severe)	Usher syndrome type 1C (OMIM 276904)
<i>USH1G</i>	Usher syndrome 1G (autosomal recessive)	Usher syndrome type 1G (OMIM 606943)
<i>USH2A</i>	Usher syndrome 2A (autosomal recessive, mild)	Retinitis pigmentosa 39 (OMIM 613809), Usher syndrome type 2A (OMIM 276901)
<i>BBS8</i>	Bardet-Biedl syndrome 8	Bardet-Biedl syndrome 8 (OMIM 615985)
<i>SPATA7</i>	Spermatogenesis associated 7	Leber congenital amaurosis 3 (OMIM 604232), Juvenile retinitis pigmentosa (OMIM 604232)
<i>COH1</i>	Cohen syndrome	Cohen syndrome (OMIM 216550)
<i>KLHL7</i>	Kelch-like 7	Retinitis pigmentosa 42 (OMIM 612943)

2.2.8.3 Next-generation sequencing

Two different targeted next-generation sequencing methods were used in this thesis, the Roche 454 GS FLX, and the Illumina HiSeq 2000 platforms. The principles of each sequencing technology were introduced in sections 1.5.1 p. 28 and 1.5.2 p. 32 respectively.

The methods that were used for Roche 454 GS FLX sequencing (in combination with NimbleGen sequence capture array) are described in detail in an appendix in section 9.1 p. 294. A brief summary is described in section 4.2.1 p. 126. The methods that were used for Illumina HiSeq 2000 sequencing are described in section 3.2.1 p. 99.

In each case these next-generation sequencing methods were used in combination with a sequence capture and enrichment method which was designed by Dr Morag Shanks (section 2.2.8.2 p. 90). Each of these sequence capture and enrichment methods used a capture design that targeted the 74 genes listed in **Table 2.4**.

A third type of NGS method used was whole genome sequencing using unchained base reads on self-assembling DNA nanoarrays, which was performed by Complete Genomics on samples that were sent to them, as described in section 6.2.2 p. 221.

CHAPTER 3

Comparison of Next-generation Sequencing with Sanger sequencing

3.1 Introduction

At the time the work in this thesis began next-generation sequencing (NGS) technologies had only been available to researchers for a few years and their use in clinical diagnostics had not yet been explored. Molecular testing for genetic conditions was instead performed by Sanger sequencing and other techniques (introduced in section 1.7.1 p.51).

The accuracy of Sanger sequencing has improved over the years since its first introduction. A review by Richterich in 1998 analysed data produced from six different sequencing projects, each of which used different combinations of sequencing machine, enzymes (e.g. Sequenase, Taq, KlenTaqTR, AmpliTaq FS), as well as either dye primers or dye terminator chemistries that were available at the time. Typical Sanger sequence data can provide a few hundred bases of high quality sequence, and the data from two of the projects reviewed by Richterich both gave less than a 1% error rate between the 70th and 330th bases sequenced (Richterich 1998).

The accuracy of Sanger sequencing can be affected by several factors, such as which molecules are fluorescently labelled, as well as the type of enzymes and sequencers used (Richterich 1998). Overall though the accuracy of Sanger sequencing is very good, and it is commonly referred to as “the gold standard” for sequencing projects.

Several different methods were used to assess the accuracy of NGS. The first method used in this thesis was to determine how NGS platforms compared with Sanger sequencing in terms of accuracy, by evaluating how well each technology identified known variants in patient DNA samples and identifying the reasons for missing variants.

Two of the leading NGS technologies available at the time this work commenced were the Roche 454 and the Illumina HiSeq (introduced in section 1.3 p. 28). However, there was very little data on the accuracy of either of these technologies, although 454 had released some data on the accuracy of their platform in detecting homopolymer tracts to attempt to allay concerns over this issue.

The second aim was to evaluate the coverage (introduced in section 1.6.2 p. 43) generated by each platform. Higher depths of coverage can increase the accuracy of sequence data, as erroneous reads will make up a lower percentage of the total reads covering a given nucleotide (Bentley *et al.* 2008).

A third method of evaluating NGS for use in clinical practice was to investigate the use of targeted capture which is performed prior to sequencing (introduced in section 1.5 p. 50). Three of the leading capture technologies at the time this work commenced were the NimbleGen sequence capture array, RainDance, and Agilent SureSelect. Each one of these technologies was also evaluated to determine its efficiency by calculating the alignment of sequenced reads generated, i.e. the percentage of sequence reads that map back to the target design as described in 1.8.1 p. 43.

Three combinations of capture technology and sequencing platforms were selected:

- RainDance microfluidic based PCR amplification & 454 pyrosequencing.
- NimbleGen sequence capture array & 454 pyrosequencing.
- Agilent SureSelect hybridisation capture & Illumina sequencing by synthesis.

These combinations were used because of compatibility issues, primarily differences in the library adaptors which prevented straightforward analysis of RainDance and NimbleGen data when used with the Illumina platform.

The cohort of patients that were selected to evaluate the technologies had been identified by Prof. Susan Downes in the Oxford Eye Hospital and had a clinical phenotype compatible with mutations in the *ABCA4* gene (Allikmets *et al.* 1997b).

The DNA samples from the patients were made available by the ORGL diagnostic service for genetic testing of the *ABCA4* gene (see methods section 2.2.1 p. 75).

The patients had previously been tested for mutations in the coding regions (and flanking sequence) of the *ABCA4* gene using Sanger sequencing. These data were made available for

the project to use as a comparison, in order to evaluate targeted capture with NGS compared to Sanger sequencing. The results of these experiments were used to assess whether these new technologies could be introduced into the clinical molecular diagnostic service of the NHS.

The main aims of the work presented in this chapter were:

- To determine whether the sequence capture and NGS technologies could identify all the variants that Sanger sequencing had identified in the *ABCA4* gene in these patients.
- To evaluate the accuracy and reliability of the NGS technologies by determining their detection rate of variants identified by Sanger sequencing, and the depth of coverage they provided.
- To evaluate the efficiency of three different sequence capture technologies by determining the alignment of sequenced reads to the target region of interest.
- To evaluate other factors that could influence the use of these technologies in a clinical diagnostic setting, such as ease of use, time spent performing experiments, and time taken for a sequencing run to be performed.

3.2 Materials and Methods

The experiments described in this chapter were performed on 16 DNA samples which came from a subset of patients with the cohort diagnosed with a macular dystrophy whose phenotype was compatible with mutations in *ABCA4*. DNA samples were extracted from whole blood by the ORGL as described in section 2.2.2 p. 75.

All 16 of the patients were sequenced using the RainDance Thunderstorm capture and Roche 454 GS FLX sequencer with Titanium reagents by Morag Shanks (University of Oxford).

All 16 of the patients were sequenced using the NimbleGen sequence capture array and Roche 454 GS FLX sequencer with Titanium reagents by Morag Shanks (University of Oxford).

All 16 of the patients were also sequenced using SureSelect capture and sequencing on the Illumina HiSeq 2000. This was performed as a collaboration with Morag Shanks (University of Oxford), and the high throughput sequencing group at the Wellcome Trust Centre for Human Genetics (University of Oxford). DNA library preparations were performed as described below.

3.2.1 Illumina sequencing

Illumina sequencing differs from Roche 454 pyrosequencing in that each cluster on a flowcell must be stimulated by a laser individually after each round of chemistry, as opposed to the 454 where each nucleotide incorporation emits a flash of light which is detected at the same moment. This leads to far greater run times of several days for Illumina sequencing as opposed to several hours for the Roche 454. The Illumina sequencing method is much cheaper however, and involves less hands-on time for preparation of samples.

The Illumina HiSeq2000 allows two samples to be sequenced at the same time. Whilst one sample is being imaged by the machine, the other sample undergoes the chemistry to prepare it for the next round of imaging. This way the imaging process (the bottleneck in the sequencing stage) is performed continuously, increasing sequencing throughput.

Illumina sequencing was performed according to manufacturer's protocols as described below, and as introduced in section 1.3.2 p. 32.

3.2.1.1 Covaris DNA fragmentation

The Covaris S220 Focussed-Ultrasonicator was switched on and the water tank filled with MilliQ water. The device was set to 20°C and left for at least 45 minutes for the water to reach temperature. Genomic DNA was quantitated to 25ng/μl, and 120μl sample volumes were aliquoted into snap cap adaptive focussed acoustics (AFA) microTUBEs. The tubes were inserted into the Covaris which was run at 50W peak incident power for 60 seconds with a duty factor of 5% and 200 cycles per burst for each sample.

3.2.1.2 End repair

The following reaction mix was prepared on ice:

Molecular biology grade water	45µl
Sonicated DNA sample	30µl
T4 DNA Ligase buffer with 10mM ATP	10µl
10mM dNTP mix	4µl
T4 DNA polymerase*	5µl
Klenow enzyme*	1µl
T4 PNK*	5µl
Total	100µl

* = Concentration of these reagents not provided by manufacturer.

This mixture was mixed by pipetting and heated at 20°C for 30 minutes. The reaction mixture was then purified using a QIAquick PCR Purification kit (see section 2.2.7.1 p. 86), and was eluted into 32µl of buffer EB.

3.2.1.3 3' end adenylation

The blunt ends of the repaired fragments were adenylated to allow adapter ligation to occur.

This was done by preparing the following reaction mixture on ice:

Eluted DNA sample	32µl
Klenow buffer*	5µl
dATP (1mM)	10µl
Klenow Exo*	3µl
Total	50µl

* = Concentration of these reagents not provided by manufacturer.

The reaction mixture was incubated at 37°C for 30 minutes, and the mixture was then purified using a QIAquick PCR Purification kit (see section 2.2.7.1 p. 86), and eluted into 10µl of buffer EB.

3.2.1.4 Adapter ligation

Sequencing adapters with a single thymine 3' overhang were ligated onto the single adenine 3' overhang of the DNA fragments. This was done by preparing the following reaction mixture on ice:

DNA sample	10µl
DNA Ligase Buffer 2x*	25 µl
PE Adapter Oligonucleotide Mix*	10 µl
DNA Ligase*	5 µl
Total	50µl

* = Concentration of these reagents not provided by manufacturer.

The mixture was then incubated at 20°C for 15 minutes, and the mixture was then purified using a QIAquick PCR Purification kit (see section 2.2.7.1 p. 86), and eluted into 30µl of buffer EB.

3.2.1.5 Preparation of adaptor ligated libraries and Illumina sequencing

Adaptor ligated DNA libraries were delivered to the high-throughput sequencing group at the Wellcome Trust Centre for Human Genetics (University of Oxford), who performed the remaining steps of library preparation, cluster generation, and sequencing. These steps involved:

- Purifying the products of the adapter ligation by agarose gel electrophoresis, extraction of the band containing the libraries, and elution of the DNA.
- Adding sequences to the adapters to enable hybridisation with primers bound to the surface of the flow cell.
- Enrichment of fragments containing adapters on both ends only. Fragments without adapters, or with just an adaptor ligated to one end cannot be used to generate clusters on the flow cell.
- Further agarose gel electrophoresis purification of the PCR products to excise and elute a gel of approximately 400bp in size to give a sequencing insert size of approximately 300bp.
- Library quantification using quantitative PCR (qPCR).
- Automated cluster generation on a flowcell. Single DNA fragments bind to the surface of the flow cell at a concentration that ensures a low probability of two fragments attaching close enough together that the clusters they form will merge. Fragments then perform a bridge amplification PCR that forms clusters of clonally amplified fragments on the flow cell.
- Sequencing on an Illumina HiSeq 2500 machine. Flow cells containing clusters are submerged in all the reagents required for DNA synthesis of a complementary strand, excluding the nucleotides. A reversible dideoxy chain termination reaction takes place that allows fluorescently labelled dideoxy nucleotides to be incorporated into the complementary strand. These molecules are then excited by a laser that scans across the flowcell, and the identity of the nucleotide base can be determined by the colour of the fluorescence detected. The termination of the chain is reversed, allowing further

synthesis of the polymer, and determination of the identity of nucleotides further along the strand.

3.2.2 Data analysis

Raw data was generated by all three combinations of sequence capture and NGS platform. The raw reads were aligned using Stampy, and variants called using Platypus. The data was all analysed by Dr Richard Copley in the Bioinformatics Team at the WTCHG (University of Oxford).

Platypus has a set of filters to remove erroneous calls which are due to errors in sequencing. These filters include

“allele and strand bias, mapping quality, base quality, insert size, sequence context and posterior probability of the variant” (Rimmer *et al.* 2014).

For all the variant calling performed using Platypus in the work described in this thesis the default filter thresholds were used throughout. Platypus benefits from not removing variants based on their read frequency, allowing it to detect variants that appear at low percentages due to mosaicism (Pagnamenta *et al.* 2012).

Sequencing metrics such as alignment of sequence capture technologies, and coverage of each NGS platform, were generated for each patient and then averaged and summarised.

The variant called files generated were analysed to determine what percentage of the variants that had been identified by Sanger sequencing had been detected by each of the combinations of sequence capture and NGS technology.

3.3 Results

The first analysis performed on the data produced by NGS was to determine whether the technologies could detect all the variants identified through Sanger sequencing.

3.3.1 Overall variant detection rates in Capture/NGS compared to Sanger sequencing in 24 patients

In total 99 variants had been identified using Sanger sequencing of the exonic and flanking regions of the *ABCA4* gene in the 16 patients. These variants included pathogenic mutations as well as polymorphisms which are well recognised in *ABCA4*.

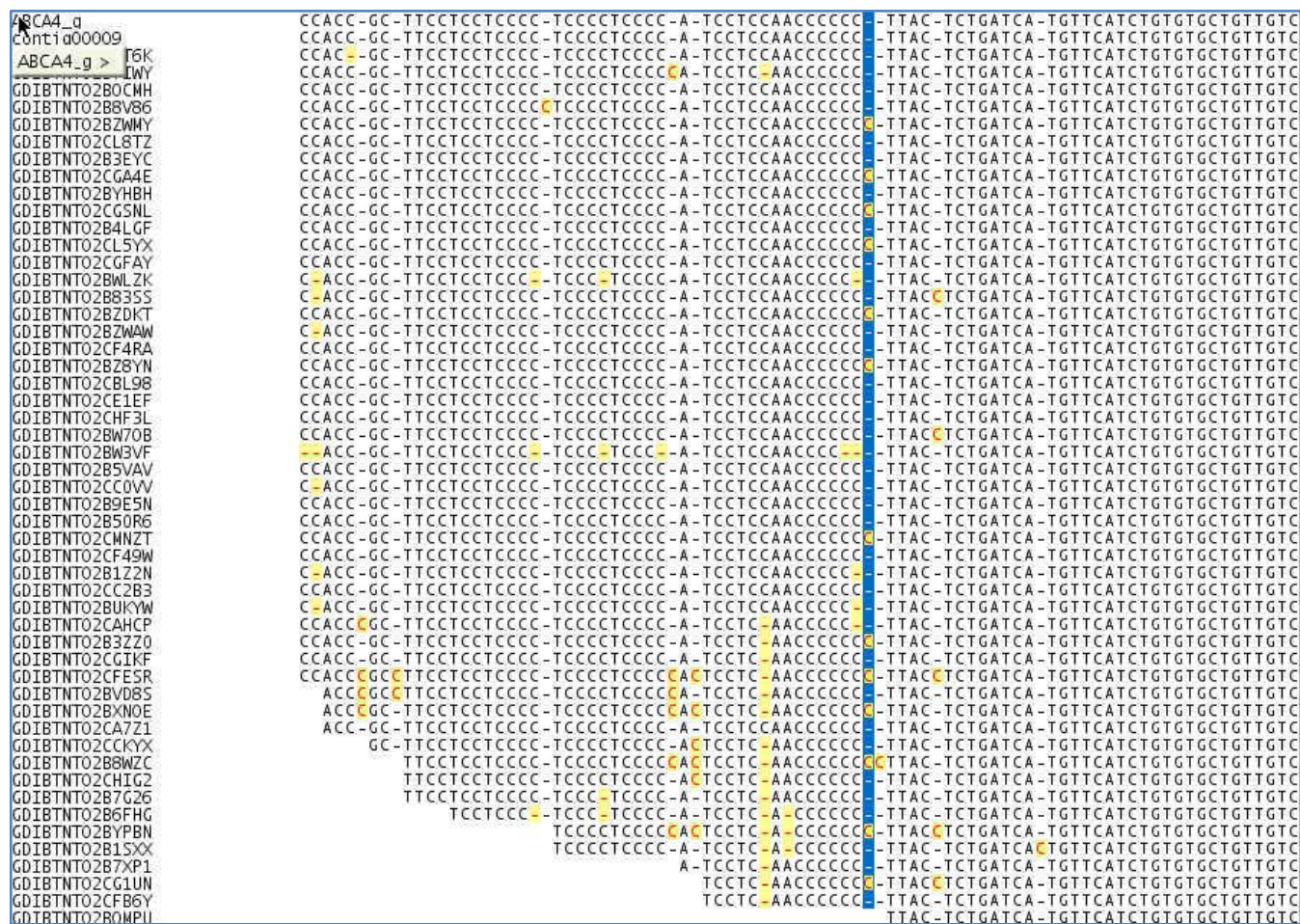
The rate of detection of both pathogenic mutations and polymorphisms were compared as the purpose of this part of the evaluation was to determine the overall ability of the technology to identify known variants. The results of this analysis are summarised in **Table 3.1**.

The SureSelect and Illumina combination showed the best ability to detect variants, with 91.9% of variants detected by the sequencing chemistry and correctly called by the software. The remaining 8.1% of control variants were not called by the software, but on closer inspection of the raw data had been detected by the sequencing chemistry and were present in the raw reads. This miscalling of variants in these positions was due to a failure of the variant calling programs which were available at the time.

The two 454 sequencing combinations performed less well, with a lower variant identification rate (75.8% and 76.8%). More variants were miscalled by the software (21.2% in each case) and a small percentage of variants were not detected by the sequencing chemistry at all. The vast majority of variants missed by the Roche 454 instrument were a single c.1356+10dupG which was present in most of the samples. This single base pair duplication is in a short homopolymer tract of G's (see **Figure 3.1**) and although the variant is detected by the pyrosequencing chemistry it was failed as a low confidence base.

Table 3.1 Overall detection rate of *ABCA4* variants.

	NimbleGen & 454	Raindance & 454	SureSelect & Illumina
Number of patients sequenced	16	16	16
Total <i>ABCA4</i> exonic variants discovered by Sanger sequencing by the NHS regional genetics lab	99	99	99
Variants identified correctly by NGS	75 (75.8%)	76 (76.8%)	91 (91.9%)
Variants mis-called by NGS	21 (21.2%)	21 (21.2%)	8 (8.1%)
Variants not detected by NGS	3 (3%)	2 (2%)	0 (0%)



Therefore, the failure to detect this is related to the software not the chemistry. Comparison data from the Raindance experiments at lower and higher coverage reveal that an increase in coverage had no effect on the detection rate of the variants in homopolymer tracts. Manual inspection of the sequencing trace revealed that the dupG is present in the raw trace but is not called by the GS Mapper Software because it is in a homopolymer tract.

Of the other variants missed by the RainDance & 454 combination:

- A pathogenic c.161G>A (p.Cys54Tyr) in exon 3 was not detected in patient 698. The average number of reads for exon 3 in this patient was 50.49. The transition from C to T leads to a run of 4 T's in the trace and again may not have been called by GS mapper because of confidence limits in homopolymer tracts.
- A benign c.6006-16G>A substitution was miscalled as a GA to TG (at bases c.6006-16 and c.6006-17) in patient 680C6. This transition was present in 9 patients, 1 homozygous and 8 heterozygous. The total number of reads for the variant called was 23/107 forward and 16/67 reverse (22% and 24% of reads respectively) suggesting that it is an artefact.
- A frameshift c.4774-16-c.4774-17delGT was missed. This is a two base pair deletion that causes a frameshift and was missed despite adequate coverage.

The missed calls in the NimbleGen sequence capture array & 454 data consisted of:

- A c.768G>T (p.Val256Val) substitution in patient 724G1 was missed due to low coverage (15x depth at that position). The variant is a benign synonymous change.
- A c.5844A>G (p.Pro1948Pro) substitution in patient 624F2, which was missed due to a very low coverage of just 8x depth at that position.
- The previously described c.1356+10dupG in a short homopolymer tract was missed in all of the patients.

In the Illumina platform no calls were missed but 6 homozygous c.1356+10dupG were called as heterozygous, likely due to an early issue with the software used.

As there was only one combination of sequence capture technology tested with the Illumina platform it is not possible to fully determine whether the higher rate of variant detection was solely due to the sequencing platform, or if this was influenced by the capture method. To determine whether this was the case, more detailed sequencing metrics were analysed, including alignment (a measure of the efficiency of sequence capture and enrichment), and coverage (a measure of the depth of sequencing throughput provided by an NGS platform).

3.3.2 Alignment of sequence capture technologies

Next-generation sequencing technologies will sequence any DNA that is included in the initial library preparation process. Depending on the extraction method used to isolate genomic DNA, this initial input can even include DNA which is not from the patient, for example DNA from parasites or pathogenic organisms.

Sequence capture and enrichment technologies are used so that only DNA of interest will be sequenced. In a diagnostic context efficient sequence capture will remove any non-human DNA sequence and only include specific sequences of interest, usually the coding sequences of genes known to have mutations causing a specific disease. The efficiency of the targeted capture step can be estimated by calculating the alignment: the percentage of sequence reads that map back to the region included in the target design ("region of interest" or "ROI").

The results of comparing alignment metrics for each of the three sequence capture technologies tested in this project are shown in **Table 3.2**.

The RainDance capture was found to have the highest alignment to the target region, with an average of 80.1% of reads mapping back to the target region of interest (ROI). The NimbleGen sequence capture array had the next highest mean of 51.6% of reads aligning to the ROI. SureSelect offered the lowest alignment with just 19.3% of reads aligning to the ROI.

The higher alignment produced by RainDance compared to the other two technologies might be due to the PCR-based method of amplification this technology uses. PCR requires exact binding

of the primers to the target DNA, which prevents an amplification from occurring in regions other than the target design, leading to very high specificity of sequence enrichment. The targeted regions are then exponentially amplified by PCR, with the quantity of target DNA doubling in every reaction, reducing the percentage of remaining template genomic DNA (i.e. non-target DNA) substantially.

In contrast, the two hybridisation methods rely on non-target DNA being washed away, which will always be an inefficient process that leads to a large quantity of non-target template genomic DNA remaining. The higher alignment of the NimbleGen sequence capture array compared to SureSelect may be due to the amplification of target sequences by ligation mediated-PCR (LM-PCR) after the capture step, which also reduces the percentage of non-target DNA in the final product. SureSelect relies purely on capture of DNA (i.e. removal of non-target sequence), and has no enrichment step to amplify target sequences.

3.3.3 Other factors affecting selection of a sequence capture technology

The percentage alignment of reads sequenced is a key factor in the selection of a sequence capture and enrichment technology, but there are other factors that are also significant. Each technology tested relied on very different methodologies, as introduced in section 1.5 p. 39.

The NimbleGen sequence capture array allowed multiplexing of samples, which meant that up to eight samples at a time could be loaded onto a single array and captured simultaneously. This increased the speed at which capture steps could be performed, but it also meant that multiple samples would have to be sequenced simultaneously, and the individual reads separated later through the use of molecular identifier (MID) tags.

Table 3.2 Alignment of captured DNA reads to target design.

	NimbleGen and 454	Raindance and 454	SureSelect and Illumina
Average alignment to ROI (%)	51.6	80.1	19.3
Range of alignment to ROI (%)	19.4 - 65.7	70.2 - 85.7	12 - 25.6

The RainDance system required primers to be synthesised by the manufacturer who was located in the USA, as well as a specialised microfluidics robot to perform the capture, and so until this was obtained, samples had to be sent to the manufacturer to be captured. As the capture step took place after the library preparation stage of the workflow, and before the clonal amplification step, this added a great deal of time and expense to the overall workflow.

SureSelect was a solution capture method, which allowed the process to be performed using standard Eppendorf tubes, and without the need to purchase expensive equipment required for each of the other two methods, which made this the quickest and easiest method to use. The very low level of alignment shown in the results for this method suggested that this technology was not an ideal choice, and so it was not for the remaining experiments. However, SureSelect was subsequently used with success in other projects, so this result may not have been representative.

3.3.4 Coverage of NGS platforms

Coverage, or depth of coverage, refers to the number of reads covering a specific locus in a genomic sequence, and was introduced in section 1.6.2 p. 43. Coverage is largely determined by the throughput of an NGS platform, although if used in conjunction with a sequence capture and enrichment technology, coverage will be influenced by the efficiency of the capture method.

The coverage provided by each combination of NGS and sequence capture method over the total region of the capture design was determined, and the results summarised in **Table 3.3**.

Overall the results showed that the Illumina platform provided a far higher depth of coverage in comparison to the 454. Illumina had an average of 847 reads covering each base, although there were still some nucleotides which were not covered at even 1x depth of coverage. In comparison 454 had a mean coverage of just 51.5 or 81.3 reads per base, and again did not cover some bases in the target design by even a single read. The 454 data also had only 59% or 86% of bases covered at the 30x minimum depth of coverage needed to accurately call

heterozygous variants (Bentley *et al.* 2008). The higher coverage of the Illumina platform was in large part caused by its much higher throughput, or number of sequences generated per run, compared to the 454.

When analysing the two different technology combinations that used 454 sequencing, the higher coverage shown by the RainDance combination is likely indicative of the higher alignment of this technology compared to the NimbleGen sequence capture array. This has resulted in more of the sequencing throughput of the NGS platform being targeted to the capture design, and as a result a higher depth of coverage over these nucleotides.

Sequencing projects will have a desired threshold at which all bases within the target region of interest (ROI) will need to be covered for the data to be reliable enough to accurately call variants, especially heterozygous variants which will only appear in 50% of reads. A small percentage of errors will appear in the data produced by any sequencing chemistry, so if coverage is too low it may not be possible to tell whether a variant is accurately called and appears at 50% frequency, or if the variant is actually a sequencing error. In some cases the genome of an individual is not the same in every cell due to somatic mutations that have occurred in a past cell from which a group of cells are descended. In such cases a very small percentage of cells in the tissue type might have a different genotype, and to detect this accurately requires even greater levels of coverage than is needed to accurately ascertain if a mutation is heterozygous and occurring in 50% of cells.

Table 3.3 Average coverage summary over entire capture design for *ABCA4* across all patients for the three sequence capture & NGS combinations

Capture technology & sequencing platform	Number of reads:			Percentage of bases covered:			
	mean	min.	max	5x	10x	20x	30x
NimbleGen & 454	51.5	0	372.7	85.0	80.1	68.8	59.0
Raindance & 454	81.3	0	1188	97.2	96.4	92.5	86.0
SureSelect & Illumina	847.0	0	3974	96.0	95.0	93.7	92.8

3.3.5 Comparison of “ease of experimental use” of each NGS platform

Although throughput, alignment, and coverage are all significant factors in assessing the utility of any NGS platform, there are other factors that affect the ability to use the technology for clinical diagnostics. These include:

- The time taken to sequence samples (both in total time, and the amount of time spent hands-on performing the process).
- The ability to automate the process with robotics.
- The difficulty of performing bioinformatic processing and analysis of the data generated.

In the time since this project began, all NGS manufacturers have developed robotic automation systems that decrease the amount of time the experimenter need actually spend “hands-on” preparing and sequencing samples. At the time the comparison was made however, Illumina required far less hands-on time due to automation of the clonal amplification and sequencing steps being included as a standard part of the workflow.

The use of automation provides several benefits. It decreases the long-term costs of experiments as less hands-on time is needed to perform experiments. It also increases the reliability of the experiments, as a robot will more consistently produce the same results, and reduces the chances of experimenter error being introduced into a multi-step workflow.

The 454 in contrast required the emulsion PCR preparation and emulsion break steps to be performed manually, both of which were time consuming and difficult. The loading and preparation of the 454 sequencing machine for a sequencing run was also more time and labour intensive than loading a flowcell into an Illumina sequencer. However, the Illumina sequencing machines require the experimenter to top up reagents during the sequencing process in comparison to the 454, in which the machine could be left to run by itself once a run had been initiated.

Another factor to be considered is that in any of the currently available NGS technologies there are many steps in the process that need to be performed between the extraction of a DNA

sample, and arriving at the final sequencing data. At each one of these steps there is the chance for errors to occur, or for there to be a difference in the particular chemistry conditions in which the reactions occur. Although not obviously apparent at the time, these differences can affect the final sequence data produced, in a number of ways:

- Reducing the throughput of data generated;
- Introducing bias into the data;
- Preventing a subsequent reaction from occurring and thus making the data completely erroneous.

As next-generation sequencing technologies are performed as massively parallel reactions, each one of these errors may only affect a specific molecule, which can in some cases be removed from the final data at the point of quality filtering. In some cases quality control steps are performed during the library preparation and clonal amplification processes, so that if a large error is detected that step of the experiment can be repeated, to prevent this error from causing further problems in downstream stages of the library preparation, clonal amplification, sequencing, or data analysis processes. Data from the 454 quality control steps will be described in section 4.2.2 p. 127.

A comparison was made of some of the other factors found to be important during the preparation and sequencing of samples on both NGS platforms, the results of which are summarised in **Table 3.4**. During the course of library preparation experiments with the 454, an issue was encountered with library preparations failing to produce libraries when analysed on the Agilent 2100 Bioanalyzer. Approximately 30% of library preparations failed to produce a suitable library, leading to the need for many repeats, and an increased use of reagents, increasing the overall time and financial costs of the experiments. Illumina sequencing was significantly cheaper than 454 even at the time this comparison was made, and the cost of this technology continued to fall far faster than 454 pyrosequencing, which was one of the primary reasons that Illumina became the NGS method of choice for most research groups.

Table 3.4 Technical comparison of ease of experimental use of each NGS platform.

Technical comparison	454 GS FLX with Titanium Reagents	Illumina HiSeq2000
Time spent 'hands on'	~28 hours for 6 samples	5-6 hours for 6 samples
Time for complete process	5 days	11 days
Bioinformatic support provided by manufacturer?	Alignment and variant calling provided by GS Reference Mapper software	No
Automation possible?	Yes	Yes
Fail rate for library preps	~30%	0%
Financial cost	More per base than Illumina	Less per base than 454

3.4 Discussion

In the work presented in this chapter, three combinations of sequence capture and enrichment technology and next-generation sequencing platform were evaluated and compared to Sanger sequencing.

Each capture technology and sequencing platform was shown to have its own advantages and disadvantages. The primary criteria that determines the utility of a sequence capture technology is the percentage alignment of sequence reads to the target design. However, ease of use and the adaptability of the capture design were also significant factors.

For the next-generation sequencing platforms the primary criteria assessed were accuracy (determined primarily by coverage and variant detection rate compared to Sanger sequencing), the time taken to perform the process, the ease of use and automation, and crucially, the ability to process and analyse the data generated from the machines.

The inability to call variants correctly 100% of the time is an observation that has since been reported by other groups. In 2013 Ratan *et al.* also identified the problem of certain variants being identified by one platform but not another and suggested three possible reasons why this might occur:

1. Low coverage.
2. Low frequency of the alternate allele in sequence reads.
3. The variant is not called by the software. This could be due to:
 - Proximity to many other SNPs.
 - Proximity to a high quality indel.
 - The variant is present in a non-uniquely aligned region.
 - Issues with BWA or Newbler (the variant calling programs used)

The observations of these authors on the problems found with variant calling correlate with the results reported in this chapter (Ratan *et al.* 2013).

Using the assessments described in this chapter the Roche 454 was initially chosen for use in further experiments. Although overall the Illumina platform appeared to be more accurate and had much higher coverage, the 454 platform had longer reads, which is a significant advantage when creating an assembly as there are more bases within each read to find overlaps. Longer reads also have more chance of being able to cover the entirety of a repeat region, and therefore sequence these regions more accurately.

A more significant advantage of the 454 however, was that the manufacturer also provided a far greater level of bioinformatic support for processing the raw data produced by the machine. At the time this work was performed the 454 platform included “GS Reference Mapper” software which allowed alignment and variant calling to be performed through a graphical user interface directly from the machine to a desktop computer.

In contrast the Illumina platform provided no support for alignment and variant calling which instead had to be performed manually using custom designed programs such as Stampy and Platypus (introduced in section 1.6.3 p. 45). These programs required specialist bioinformatic support including trained bioinformaticians to design, run, and modify them for the settings needed for a specific sequencing project. This expertise was unavailable, and therefore the 454 was chosen as the NGS platform to be used for the next experiments.

The NimbleGen sequence capture array was chosen for the targeted capture as it was much cheaper, quicker, and easier to use than RainDance, and appeared to provide far higher alignment than SureSelect. It also had the option of multiplexing samples, allowing up to 12 samples to be hybridised and captured simultaneously.

In the time since these experiments were performed however, the alignment of SureSelect has improved, and this has become the most commonly used capture technology by the field of NGS as a whole, in part due to the ease of use and lack of special equipment required to perform it. Illumina have also improved both their throughput, and their data analysis software, and so this has now become the NGS platform of choice.

The selection of 454 and the NimbleGen sequence capture array prompted further experiments to investigate the ability of these technologies to identify pathogenic variants in a group of patients with retinal dystrophies and no molecular diagnosis. This included the development of a data analysis pipeline that would be able to identify pathogenic variants from the list of variants called in a patient sample. The next chapter describes the results of these investigations.

CHAPTER 4

Investigating the clinical utility of NGS in patients with Inherited Retinal Dystrophies

4.1 Introduction

The work in Chapter 3 described the investigations used to determine which combination of sequence capture and next-generation sequencing (NGS) platform to use for preliminary investigations into the clinical utility of NGS. These decisions were made based on experiments performed to detect known variants which had already been identified by Sanger sequencing.

In the work described in this chapter the NimbleGen sequence capture array was combined with the Roche 454 sequencing platform, to sequence and analyse DNA samples from a cohort of patients with inherited retinal dystrophies (IRD) in whom no previous genetic diagnosis had been made.

This chapter falls into three sections:

- The first discusses quality control experiments that are performed at each part of the library preparation, emulsion PCR, and sequencing stages of the targeted next-generation sequencing method used.
- The second part describes the assessment using this method of a cohort of patients with IRD in whom no molecular diagnosis had been made. This included the development of a bioinformatic pipeline to analyse and filter the raw data produced from sequencing.
- The third part describes the assessment of a single patient using exome sequencing.

The sequencing described in this chapter was performed on a cohort of 55 patients who were seen at the Oxford Eye Hospital by Dr Susan Downes and who had a clinical diagnosis of maculopathy, cone-rod dystrophy, or retinitis pigmentosa (see section 2.2.1 p. 75). DNA from these patients was captured using a NimbleGen sequence capture array designed by Morag Shanks to target 74 known retinal disease genes (see **Table 2.3** in section 2.2.8.2 p. 90). The capture design was made by targeting all the known retinal genes on the RetNet database (available at: <https://sph.uth.edu/retnet/sum-dis.htm#A-genes>) that had been reported to have an association with retinal dystrophies at the time these investigations were started. Following

the capture step, the DNA was sequenced on the 454 GS FLX platform with Titanium reagents, as described in section 9.1 p. 294.

The cohort of 55 patients included 14 patients diagnosed with a maculopathy in whom a secure molecular diagnosis had already been achieved using Sanger sequencing. These 14 patients were used as positive controls to ensure that the capture, sequencing and bioinformatics pipeline was as robust as had been found in the validation experiments described in Chapter 3. The remaining 41 patients had no known genetic diagnosis.

The 454 sequencing machine produces raw data in standard flowgram format (SFF) files. These data were then assembled into a continuous sequence, and a list of variants called, using 454 GS Reference Mapper software.

A single run of the 454 machine in this project produced approximately 4000 variants across twelve multiplexed patient samples. For Mendelian conditions only one or two pathogenic variants are needed to cause a disease phenotype, so a minimum of just 12-24 of the 4000 variants in the data from twelve patients would be expected to be pathogenic. This number makes up less than 1% of the total number of variants called from a sequencing run.

At the time these investigations began very few bioinformatic pipelines had been developed. The first aim of the work described here was to create a novel bioinformatic data analysis pipeline. This would filter out the benign variants, leaving only a handful of potentially pathogenic mutations that could then be examined in further detail as plausible causes of the condition in the patient.

The aims of the investigations described in this chapter were to:

1. Obtain experimental quality control (QC) data for the library preparation stage of the 454 protocol to ensure that each step was optimised before proceeding to the next step.
2. Analyse the quality of the sequencing data produced and understand potential biases and sources of error in the final sequences on individual patients.
3. Develop a data analysis pipeline capable of identifying pathogenic variants from a variant called file (VCF) generated by NGS of a patient DNA sample.
4. Identify the pathogenic mutations responsible for the disease in a cohort of IRD patients without a molecular diagnosis.
5. Determine the clinical phenotype in those with or without a diagnosis.
6. Determine the detection rate overall and in different clinical subtypes.
7. Test the bioinformatic pipeline in unknown cases.
8. Test the ability of several bioinformatic pathogenicity prediction programs to find the causative variant in patient 525A6, who had a complex phenotype.

For the QC data obtained for aims 1 and 2 the implications of the results at each stage will be discussed as they are described.

4.2 Materials and Methods

55 patients were recruited for this investigation who had a variety of clinical phenotypes (introduced in section 1.8.1 p.54). These included:

- Maculopathies (excluding patients with Stargardt disease)
- Cone-Rod Dystrophies
- Retinitis pigmentosa

43 of the 55 patients had either cone-rod dystrophy or retinitis pigmentosa and had previously received no genetic testing. The experiments described in this chapter aimed to find a secure molecular diagnosis for these patients using targeted next-generation sequencing.

The remaining 14 of the 55 patients had been diagnosed with a maculopathy and already had a secure molecular diagnosis. These patients were used as positive controls to determine whether the methodology used could detect their pathogenic mutations.

DNA was extracted in the Oxford Regional Genetics Laboratory (ORGL) from patients' whole blood samples using Qiagen Autopure LS and EZ1 systems supplied by Qiagen (UK).

DNA samples were then used to perform sequence capture and next-generation sequencing, using a NimbleGen sequence capture array and a Roche 454 GS FLX sequencer with Titanium reagents.

4.2.1 Summary of NimbleGen sequence capture and 454 sequencing process

The 454 sequencing process was originally designed for whole genome sequencing, so when sequence capture and enrichment technologies are used they have to be inserted into the normal workflow. **Figure 2.1** is a flowchart showing the normal 454 workflow in blue. The NimbleGen sequence capture array steps have been additionally introduced in orange. Downstream data processing and analysis is shown in green. Full details of the methods used can be found in section 9.1 methods appendix p. 294.

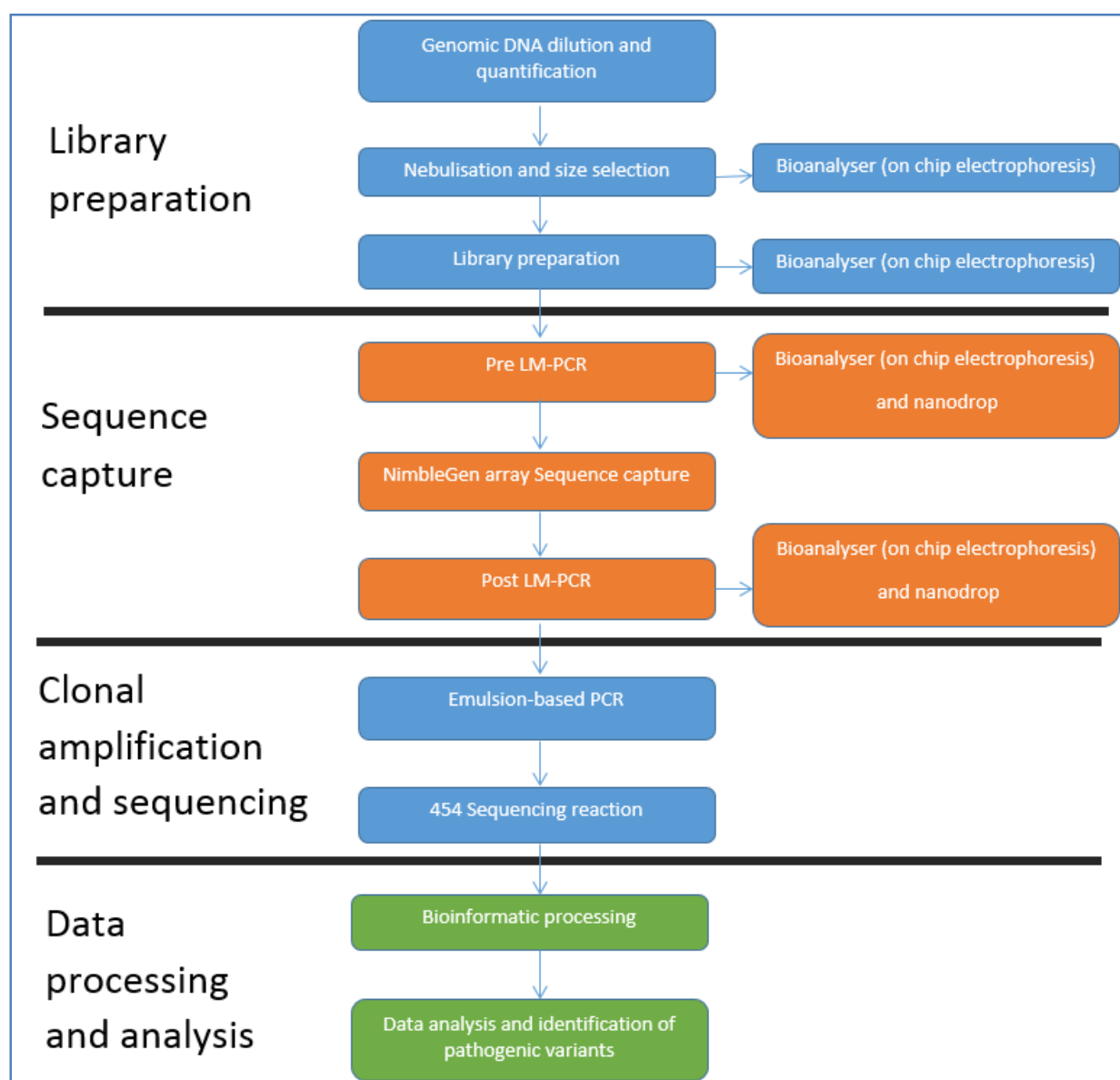


Figure 4.1 Flowchart of 454 sequencing methodology in conjunction with NimbleGen sequence capture array. The normal 454 methodology is in blue, the sequence capture steps introduced are in orange, and the data processing and analysis steps are in green. LM-PCR = Ligation mediated polymerase chain reaction.

4.2.2 Experimental quality control data

The 454 library preparation, clonal amplification and sequencing procedures are all comprised of multiple steps. At each one of these steps there is a chance for errors to occur which will ultimately affect the quality and accuracy of the final data produced. Quality control experiments are performed between these steps to identify these errors, allowing a failed process performed on a sample to be repeated before it proceeds to the next step in preparation for sequencing.

4.2.2.1 QC data for library preparations

4.2.2.1.1 QC of nebulisation and size selection

The initial step of library preparation was to fragment genomic DNA into short oligonucleotide sequences of less than 1000bp in size, as fragments larger than this do not amplify well during the emulsion PCR (emPCR) step and this ultimately leads to a higher percentage of erroneous reads, decreasing the throughput of the sequencing run. The DNA was fragmented using the process of nebulisation which uses high pressure nitrogen gas inside nebuliser cups which fragments the DNA. The DNA fragment sizes are affected by the length of time of nebulisation, which can be adjusted experimentally. Several problems can be encountered during the nebulisation step. These include inefficient or failed nebulisation, which can occur because nebulisation requires repeated calibration of the nitrogen gas pressure. If this is not performed the pressure falls, resulting in a shift of the average DNA fragment size to the right. In some cases these errors can be rectified by nebulising the sample again for a further ten seconds, otherwise the nebulised sample must be discarded and the nebulisation repeated with a new sample.

Following the nebulisation step a double Solid Phase Reversible Immobilisation (SPRI) beads reaction is then performed to remove fragments of less than 300bp in size. If double SPRI has failed to remove enough of the small fragments below 300bp they can be preferentially amplified in the emulsion emPCR step and will lead to a decrease in the overall throughput from a

sequencing run, as the shorter fragments still occupy the same number of wells in the Pico-Titer Plate, but produce less data. These stages are described in detail in Appendix sections 9.1.1.1 to 9.1.1.3 pp 295-295.

To determine whether nebulisation and double SPRI produced DNA fragments within the optimum size range, the DNA was sized on an Agilent 2100 Bioanalyzer using a DNA 7500 chip (described in section 9.1.1.4 p. 295). Successful nebulisations showed a bell curve between 300bp and 1000bp with a peak close to 500bp. A representative example of these results is shown in **Figure 4.2 A**.

Figure 4.2 B shows a sample with a higher than ideal percentage of fragments over 1000bp in length with a shift to the right. However, in this example enough of the fragments were within the 300-1000bp range and the sample was still used for emulsion PCR and sequencing without re-nebulisation.

Figure 4.2 C shows an example of nebulisation failure. This resulted in a flat trace. There are several explanations for this, such as the nebuliser cup not being tightened securely, or a crack being present in the cup which prevents it from being airtight, and so preventing nebulisation from occurring. In this case the sample needed to be re-nebulised

The SPRI step was efficient and no problems were encountered with an excess of smaller fragments (in which the curve would have been shifted to the left).

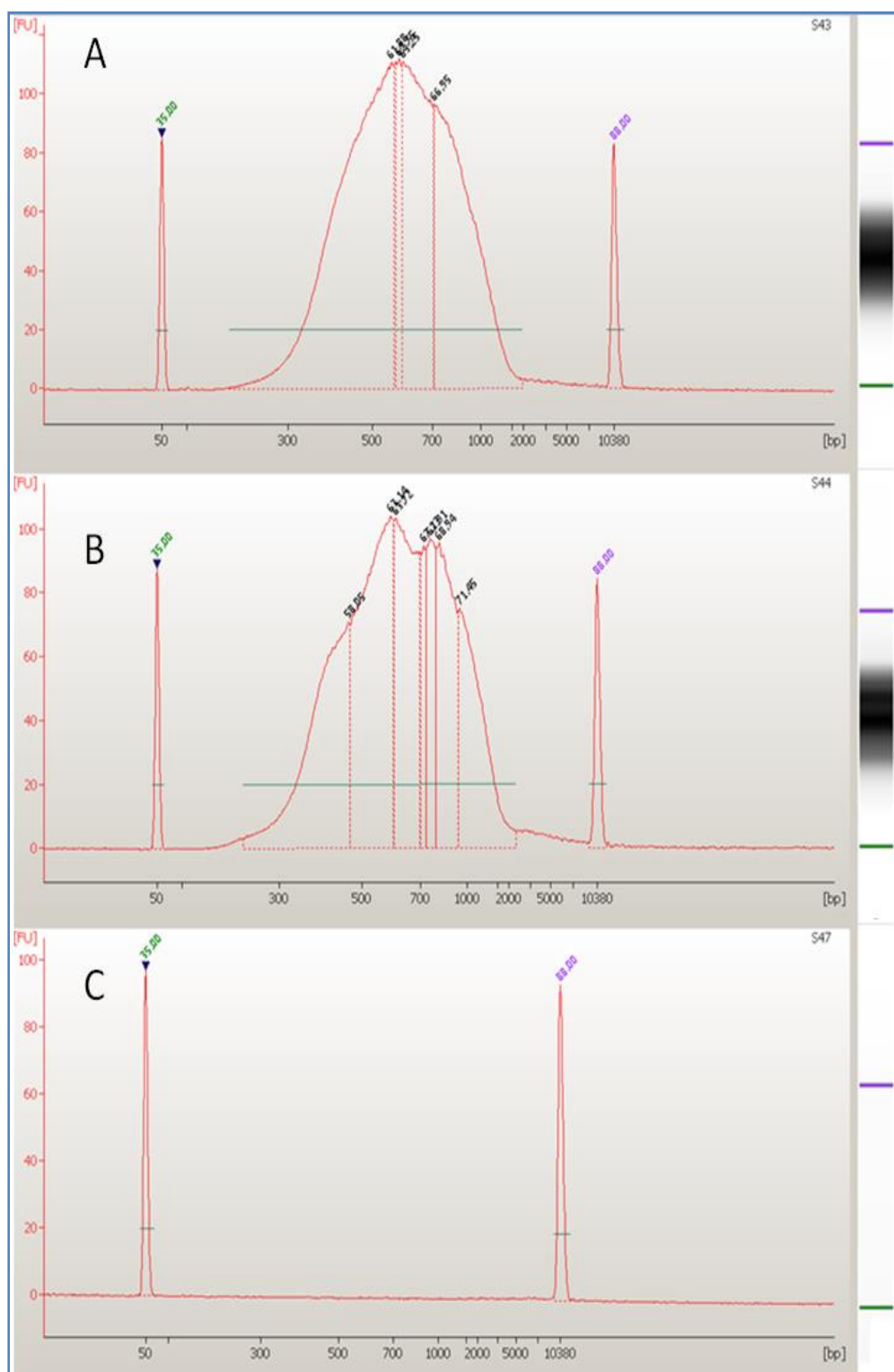


Figure 4.2 DNA nebulisation quality control data. Examples of Agilent 2100 Bioanalyzer traces run on nebulised DNA using a DNA 7500 chip. **A)** An optimal nebulisation with the majority of fragments between 300bp and 100bp. **B)** A less optimal nebulisation with the peak shifted slightly to the right. **C)** A failed nebulisation with no peak visible.

4.2.2.1.2 Library preparation QC

The next step after nebulisation is the library preparation in which the sample DNA fragments are end polished, adaptors are ligated, and the libraries immobilised into single stranded DNA (described in sections 9.1.1.5 to 9.1.1.11 pp 296-298). Because multiple purification steps are involved it is possible for significant quantities of DNA to be lost. Therefore, at the end of the library preparation the amount of single stranded DNA is quantified using an RNA Pico 6000 chip on the Agilent 2100 Bioanalyzer.

Figure 4.3 A shows a representative example of a successful library preparation with a small size marker peak on the left, and a much larger peak indicating the single stranded library preparation next to it. **Figure 4.3 B** shows a failed library in which the quantity of DNA is too low. The marker peak here is much larger than the DNA peak. **Figure 4.3 C & D** show examples of failed libraries where all the DNA has been removed.

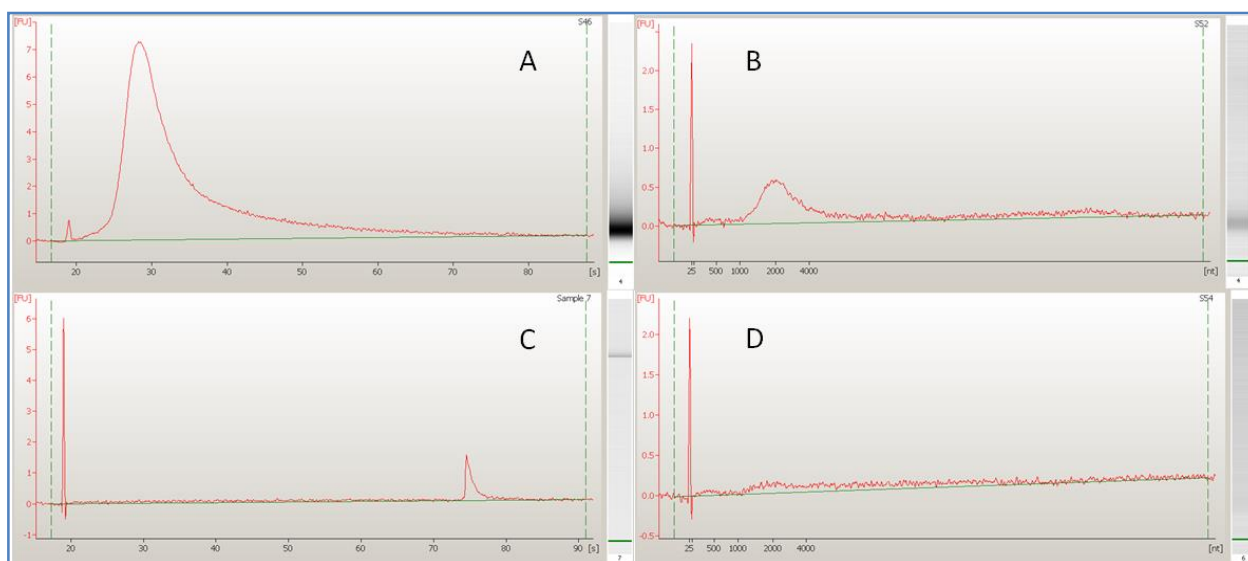


Figure 4.3 Quality control data for DNA library preparations. Examples of Agilent 2100 Bioanalyzer traces ran on DNA library preparations using an RNA Pico 6000 chip and reagents on **A)** A successful library preparation with a high quantity of DNA present. **B)** A failed library preparation where only a very small amount of DNA is present. **C)** & **D)** Failed library preparations with no DNA.

4.2.2.1.3 QC of Pre-capture and post-capture LM-PCR

Prior to performing the capture step a Ligation Mediated PCR (LM-PCR) was performed as described in section 9.1.2.1 p. 299. The samples were checked again on the Agilent 2100 Bioanalyzer to confirm that the LM-PCR steps were successful. **Figure 4.4 A** shows a failed library preparation in which a small primer dimer peak is visible (shown by red arrow). This shows an error occurred in the LM-PCR step, which had to be repeated. **Figure 4.4 B, C & D** all show examples of successfully amplified libraries. Despite having different peak heights there was sufficient DNA within the 300-1000bp reads in each library to proceed with capture. **Figure 4.4 C** was the most optimal of the three with a higher curve. Each sample was also analysed by nanodrop spectrophotometry to ensure DNA purity. Successful samples had an A_{260}/A_{280} value between 1.7 - 2.0, and a yield of >1.0 μg upon nanodrop spectrophotometry.

The results of these QC data indicated that the DNA was of sufficient quality, size and quantity to proceed to the sequencing steps.

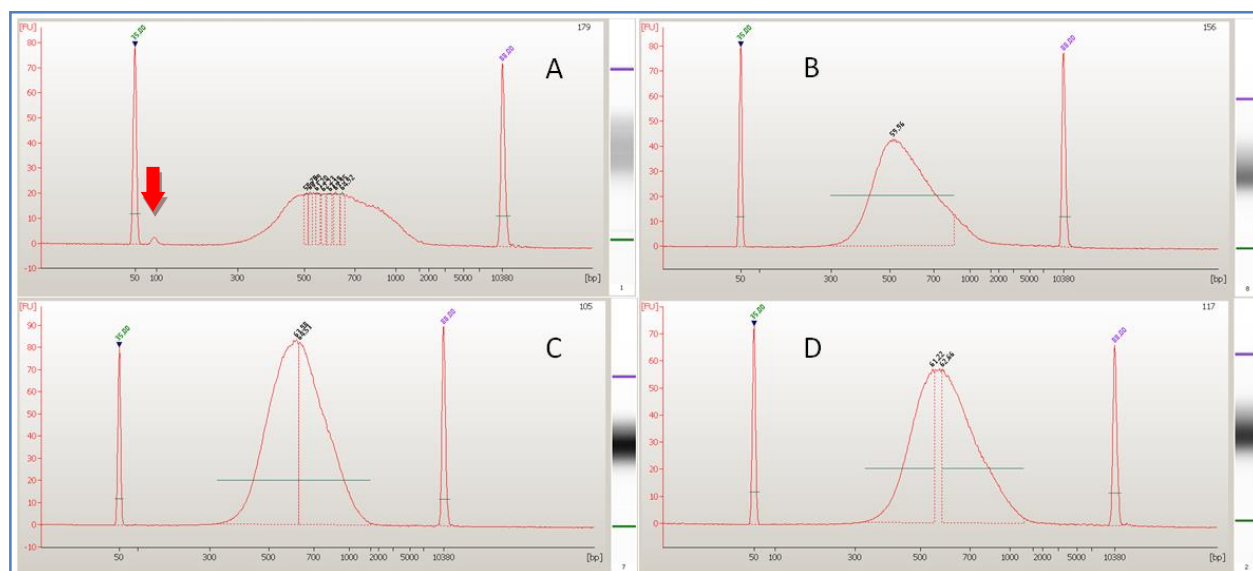


Figure 4.4 Quality control data for pre-capture LM-PCR amplified DNA libraries. Examples of Agilent 2100 Bioanalyzer traces using DNA 7500 chips and reagents ran on post LM-PCR, pre-capture libraries **A)** A failed amplification showing a small peak near the lower marker, indicated by the red arrow. This is due to the presence of primer dimers. **B), C), and D)** Successful amplifications generated from LM-PCR.

4.2.2.2 QC following sequencing

All next-generation sequencing chemistries and the associated detection methods have errors which could significantly affect the quality of the output data. In addition to the possible errors in library preparation just mentioned, errors can also be introduced in the following steps:

- Emulsion PCR errors:
 - More than one fragment of DNA attaching to a bead prior to amplification.
 - No fragments attaching to a bead prior to amplification.
- Sequencing errors
 - The inaccuracy of detecting homopolymer tracts in pyrosequencing.
 - Too many packing beads in a well.
 - Failure of nucleotide incorporation.

For this reason control DNA samples are included in each sequencing run. These are provided by the manufacturer and are DNAs which have gone through the entire size selection, library preparation and clonal amplification stages.

The 454 sequencing platform uses sophisticated software, GS Run Browser at the end of each sequencing run to identify potential sources of sequence error and provide output metrics to assess the quality of the data produced.

4.2.2.2.1 Number of wells passing quality filters

For each sequencing run the software displays images showing the wells in which reads passed the filter in green (**Figure 4.5 A**), wells in which there had been a sequencing error in red (**Figure 4.5 B**), wells that contained control DNA in blue (**Figure 4.5 C**), and a composite image of all three (**Figure 4.5 D**).

If such an image were to show blue wells but no green wells it would indicate that there has been a significant error in the library preparation or emulsion PCR steps, but a successful sequencing run. If there are neither blue nor green wells present, then it would indicate that an error has taken place during the sequencing run itself.

4.2.2.2.2 Mechanical errors

The white arrow in the image in **Figure 4.4** indicates a dark area on all the images. This is indicative of a loading error. In this case an air bubble formed in the PTP device when the plate was being loaded, preventing any beads or reagents from being deposited into these wells. As a result, no sequencing reactions occurred in this area, reducing the amount of data produced. This particular error can be prevented by tilting the plate when pipetting in the different layers of PPIase and enzyme beads (see section 9.1.4.5 p. 311).

The GS Run Browser software also shows this information quantitatively in a bar chart as shown in **Figure 4.6**, where the reads that have failed quality controls can be visualised quantitatively. The fail column is made up of two types of error, dot and mixed. Any reads in which more than 5% of the time there are three successive flows without a nucleotide incorporation are assigned as “dot errors”. These errors are usually the result of a failure in the reaction conditions of that specific well. Reads with more than 70% of flows incorporating a read are assigned as “mixed errors”. These indicate that a bead has more than one type of fragment attached to it, rather than a clonal amplification of multiple copies of the same fragment. In **Figure 4.6 A** the data shows a high percentage of reads passing filter and a low failure rate indicating a successful sequencing run while **Figure 4.6 B** shows a sequencing run with a significantly higher error rate.

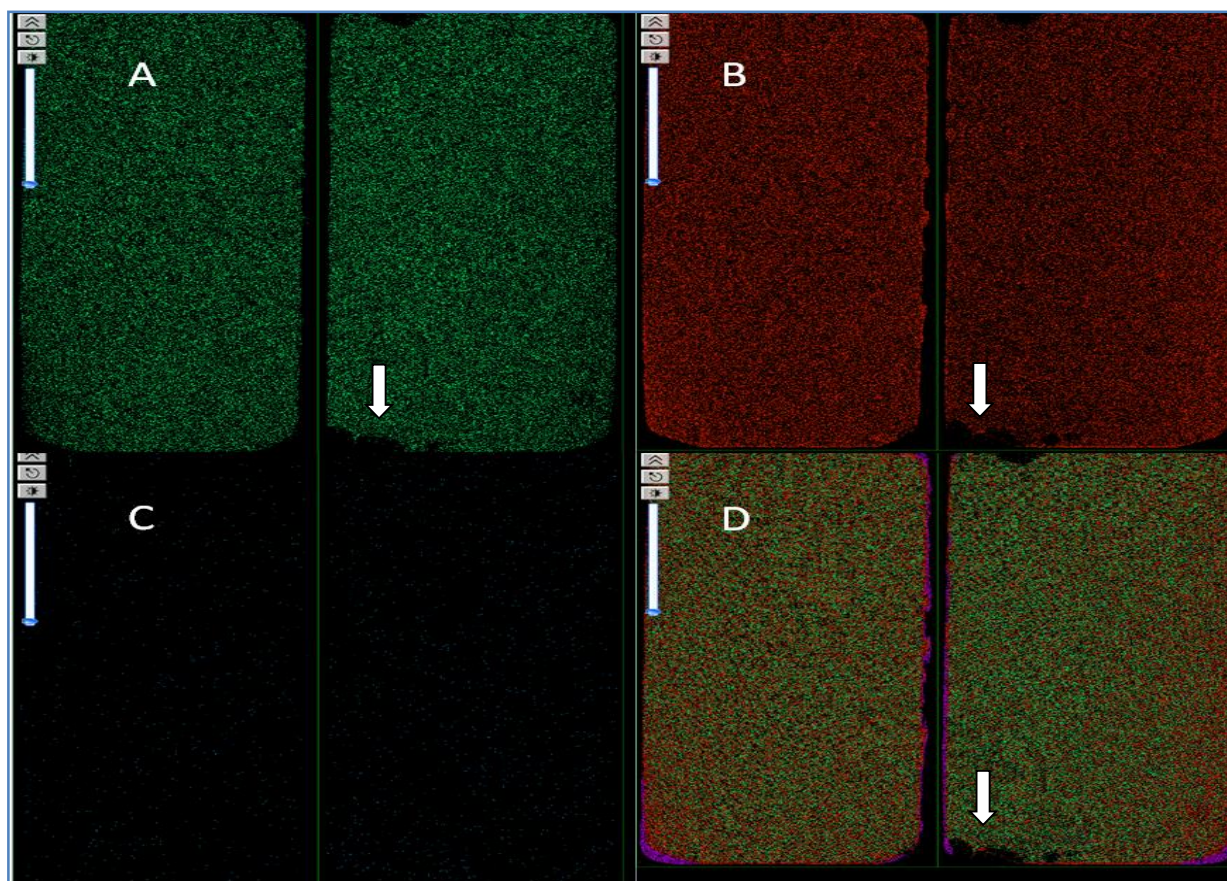


Figure 4.5 Image of wells producing data that passed quality control data filtering. GS Run Browser software images of 454 imaging data from a successful sequencing run performed on two areas of a single Pico Titer Plate (PTP). **A)** Green wells contained DNA which produced reads that passed the filter. **B)** Red wells produced data which did not pass the filter **C)** Blue wells contain control DNA. **D)** A composite image of all three. The white arrow shows a region where no data has been produced due to an air bubble forming when the DNA bead layers were being loaded into the PTP device.

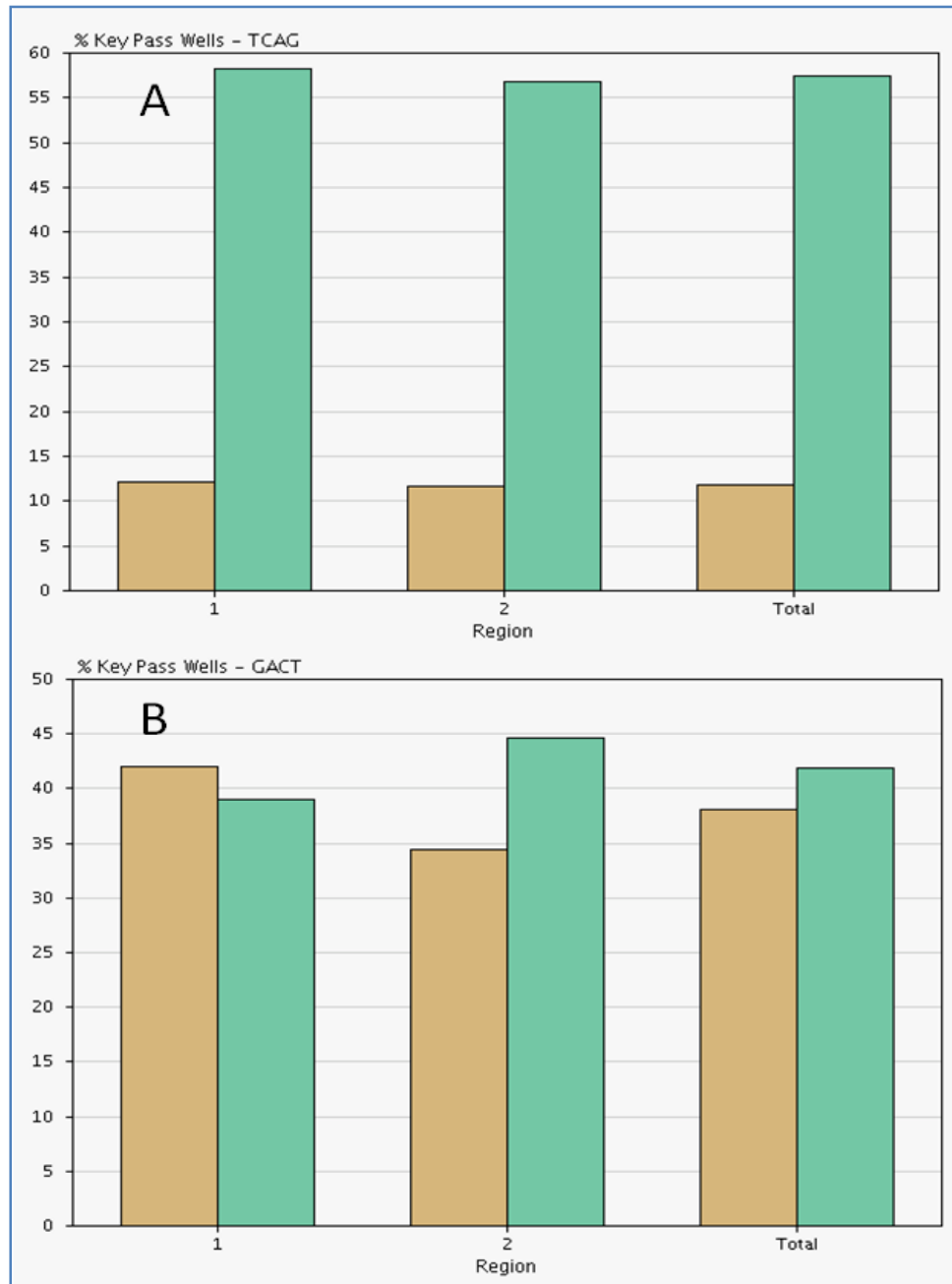


Figure 4.6 Percentages of library reads that passed quality control filtering. An example GS Run Browser software image of the percentage of library reads that passed filter (green) and the percentage that failed (beige) in two example runs. **(A)** A sequencing run in which each region has over 55% of reads passing the filter, and only 12.5% have failed, which would be within normal parameters. **(B)** A sequencing run in which only 42% of reads have passed the filter, and over 37% have failed, suggesting an error has occurred either in the sequencing run itself or in the library preparation or clonal amplification steps.

4.2.2.2.3 Chemistry errors

Homopolymer tracts (i.e. several identical nucleotides) are detected as a longer flash of light; however, there is not a linear relationship between the number of nucleotides and the light signal, as introduced in section 1.3.1 p. 28. The GS Run Browser displays the number of reads which contain a homopolymer tract, sorting them by size. In the example in **Figure 4.7** only a small percentage of the homopolymer reads are longer than 4bp, beyond which their detection becomes inaccurate and the software filters out the reads.

In the run shown in **Figure 4.7** only a small percentage of the total signals detected were over four bases of homopolymer, which suggests that the majority of the data produced should be accurately determined by the pyrosequencing.

The GS Run Browser software also provides data on the level of accuracy with which the control DNA was sequenced. Control DNA reads would be expected to have no errors in the sequence of the prepared libraries, so the error rate in these reads gives an indication as to what errors may have been due to the chemical and mechanical conditions of that particular run. An example of this data is shown in **Figure 4.8**.

Data is shown for each region individually and for an average of the two. The green column indicates the percentage of reads that have a 100% match to 200bp of control DNA reads. The beige column indicates the percentage of reads with a 98% match, and the pink column a 95% match. **Figure 4.8 A** shows better metrics than **Figure 4.8 B** as all three columns are higher, and in particular the percentage of 100% matches is higher. This metric provides an indicator of the expected error rate in library reads for that particular run due to mechanical and sequencing chemistry conditions that were unique to that particular run.

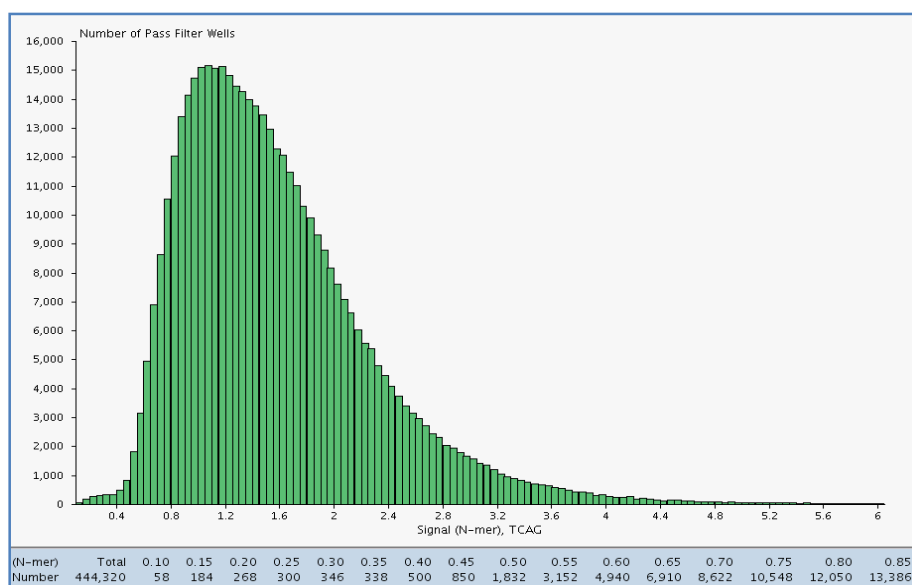


Figure 4.7 Homopolymer signals in a sequencing run. The GS Run Browser software displays the number of homopolymer signals detected of varying lengths in a sequencing run, and the frequency at which the size of each homopolymer appears. The majority of homopolymers are below the four nucleotide level at which accuracy decreases substantially.

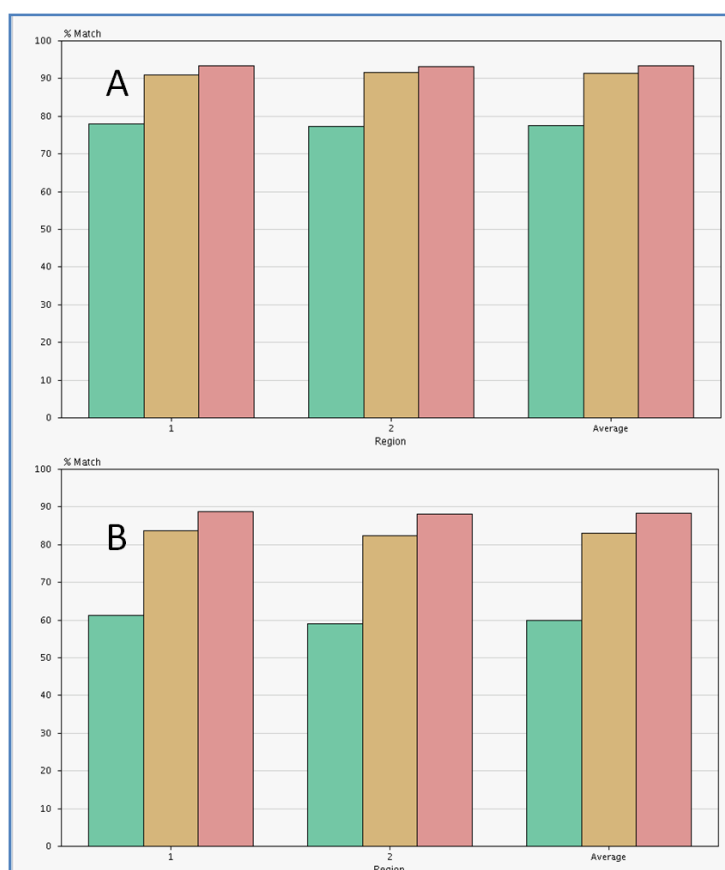


Figure 4.8 Control DNA metrics. GS Run Browser software image of the percentage of control DNA reads that match to expected sequence perfectly in green, at >98% in beige, and at >95% in pink, in two different sequencing runs. Data is shown for each of two regions individually and as an average of the two. **A** shows better metrics than **B** as all three columns are higher, and in particular the % of 100% matches are higher.

4.2.2.2.4 Sequencing metrics

In addition to the errors described above, there are numerous important metrics that needed to be assessed which indicate the success of initial library preparation steps such as nebulisation, etc.

The GS Reference Mapper software can display a graph showing the number of reads of each read length in a sample sequencing run, or from two runs combined. The length of sequence reads is largely determined by the efficiency of nebulisation at the start of the library preparation process.

The modal read length is approximately 390 base pairs long, and the majority of reads are between 200bp and 500bp in length. The longer the read length of each library the more data are produced, so an optimal run will have a low first peak below 100bp, and a high second peak as far to the right as possible. The data shown in **Figure 4.9** were typical of sequencing runs performed on the 454 platform in this project.

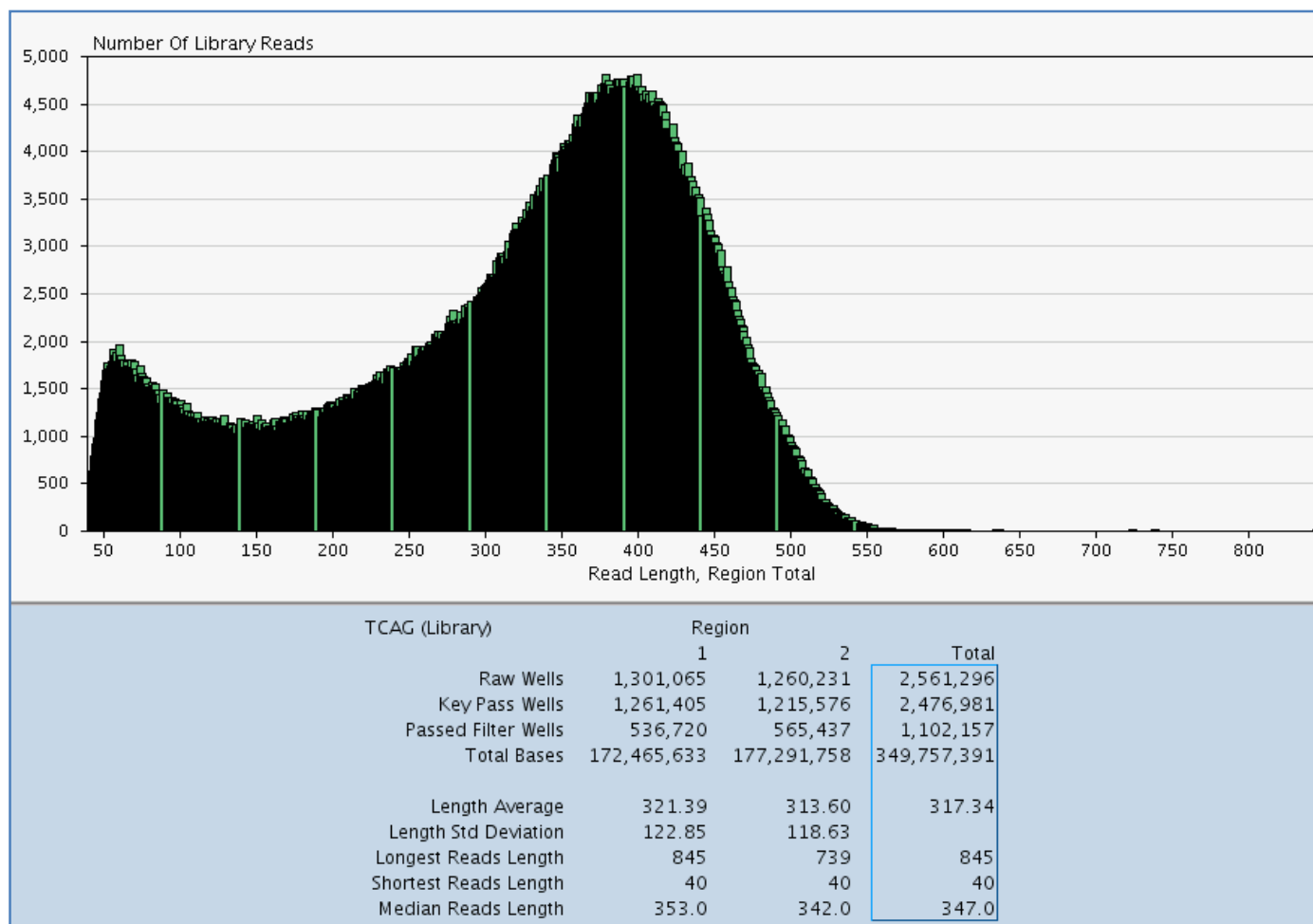


Figure 4.9 Read lengths in an average sequencing run. Example of a GS Run Browser software image of the number of library reads of different read lengths in a successful sequencing run. The modal read length is approximately 390 base pairs long, and the majority of reads are between 200bp and 500bp in length.

4.2.2.2.5 Alignment metrics and effect on variant calling

The GS Reference Mapper software provides data on alignment and variant calling. Using the GS Reference Mapper software it is possible to visualise the overlaid sequencing reads, as shown in **Figure 4.10**. Overlapping reads are layered over one another to compile a continuous sequence. By looking at this data it is possible to see where in a read a variant appears, which is important as the chemistry used in sequencing by synthesis methods is more likely to generate errors towards the ends of reads (Brockman *et al.* 2008).

In the example in **Figure 4.10** a c.3004 C>T substitution in exon 20 of *ABCA4* is highlighted in blue. In approximately half the reads visible, one of the guanine bases in the two-base homopolymer tract has been deleted, and an adenine base has appeared, suggesting that the mutation is heterozygous. To the right of the variant there are a series of homopolymer tracts. At the end of some of these reads deletions are visible, which are likely caused by an incorrect detection of the homopolymer. This is most evident on the cytosine homopolymer tract where the reference sequence shows five cytosine nucleotides in a homopolymer tract, but one cytosine has been deleted in the majority of the sequence reads. This demonstrates the low level of accuracy for detecting homopolymer tracts through pyrosequencing.

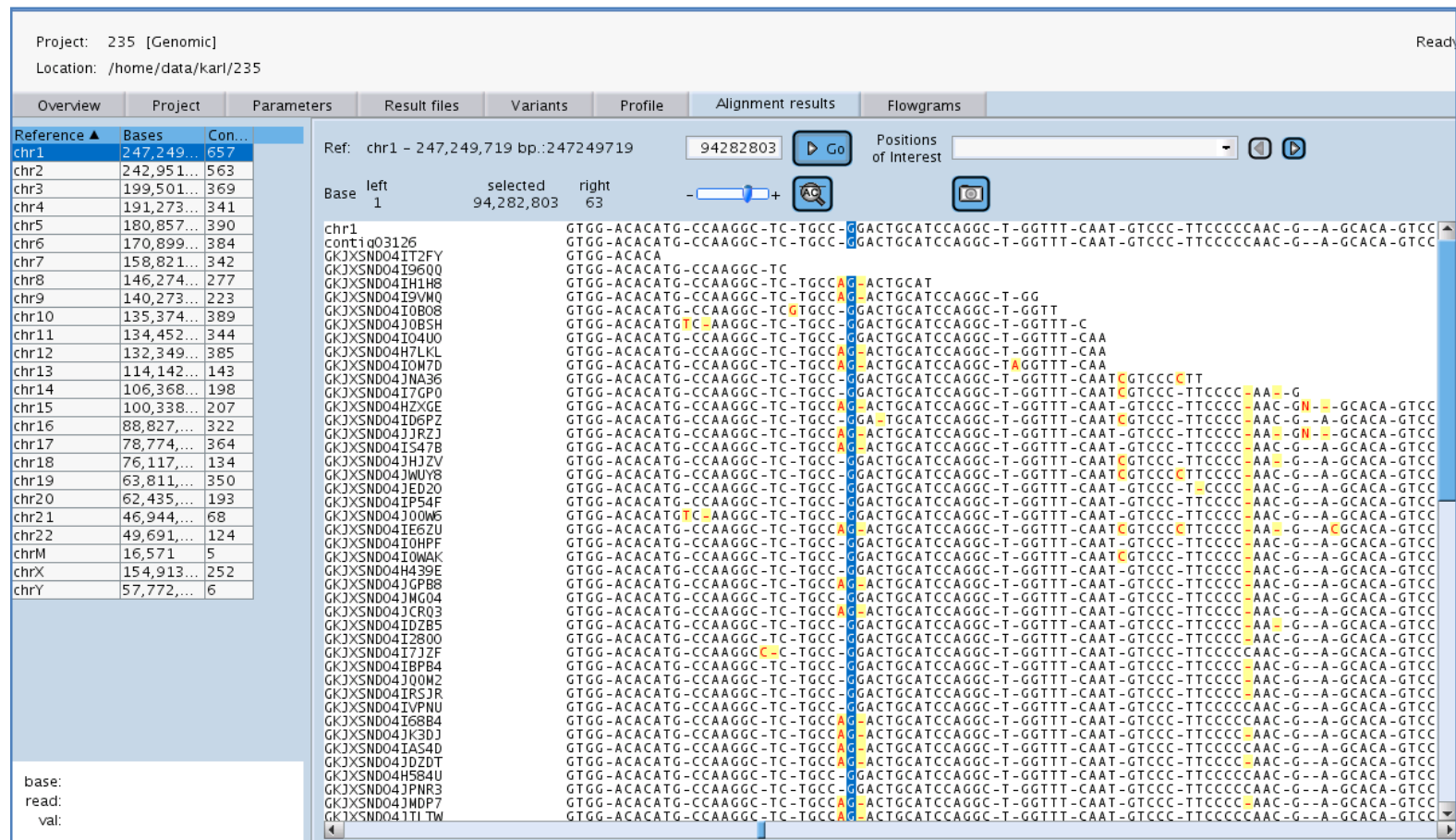


Figure 4.10 Raw data analysis using the GS Reference Mapper software. An example screenshot shows how reads are overlaid upon one another where the sequences of the reads match to create a continuous consensus sequence. Sequence variants are highlighted in red and yellow.

The GS Reference Mapper software also provides a visualisation of how variants are called, with an image of a reference sequence and the target sequence overlaid, and a third line showing the differences between the two from which substitutions, deletions and insertions can be identified. An example of this data is shown in **Figure 4.11**.

Number of Bases – Reference

Number of Bases – Read (reverse complement)

Number of Bases – Difference

Figure 4.11 Example of a flowgram on the GS Reference Mapper software. This flowgram shows the region in **Figure 4.10**. An adenine insertion is visible as a positive green bar, and a guanine deletion as a negative black bar, indicating a G>A substitution.

4.2.3 Exome sequencing

Exome sequencing was performed on patient 525A6 by the high throughput sequencing group at the Wellcome Trust Centre for Human Genetics using a SureSelect exome capture kit provided by Agilent (UK) and the Illumina HiSeq 2000 sequencer. A variant called file was generated by Richard Copley.

4.3 Results

Targeted sequencing of 74 genes associated with retinal disease was performed using NimbleGen sequence capture array and the Roche 454 GS FLX sequencer with Titanium reagents on 55 patient samples who had been diagnosed with macular dystrophy, cone-rod dystrophy, or retinitis pigmentosa.

4.3.2 Development of bioinformatic data analysis pipeline

Data generated by NGS machines requires extensive bioinformatic processing, filtering, and analysis to provide medically useful information, as previously introduced in section 1.6.4 p. 48.

The initial pipeline developed to filter the NGS data is shown in **Figure 4.12**. First, all variants with a frequency of less than 20% were removed, as these were considered likely to be sequencing errors (many of the potential sources of which were discussed in the previous sections on quality control). Real variants should instead be expected to show approximately 50% or 100% read frequency depending on whether they are heterozygous or homozygous. If a patient is homozygous for the wild type allele then they would not be expected to show any reads showing variant sequence, even at a low level, unless these were produced by sequencing errors.

Secondly, all variants were removed which were not annotated as non-synonymous single nucleotide variants (nsSNVs) i.e. a single base substitution that changes the amino acid translated by the codon it is in. This included all intronic variants, all INDELS, and all synonymous SNVs (ssSNVs). Synonymous SNVs were presumed to be benign as they had no effect on the amino acid sequence that would be translated.

The next step in the bioinformatics pipeline was to analyse the remaining variants in more detail. This was done first by using bioinformatic programs such as PolyPhen2, and SIFT (introduced in section 1.6.3.1 p. 45). The scores for these programs were obtained through another pathogenicity prediction program; Alamut Visual (available at <http://www.interactive-biosoftware.com/alamut-visual/>), which provides an interface that allows scores from a number of other prediction programs to be easily obtained.

All variants that were predicted to be benign by PolyPhen2 were removed and used to create a filter list which subsequent sequencing runs were filtered by. Finally, a literature search was performed to determine whether any of the remaining variants had been previously identified as disease causing or disease associated.

This initial analysis pipeline was then modified to create an improved pipeline shown in **Figure 4.13**. This was to ensure that a) intronic variants located close to splice sites were included and b) indels were included. The filter list generated from the initial pipeline was used to filter the subsequent raw data that was filtered using this second pipeline. Once again all variants with a frequency of 20% or less were removed for quality control. Variants were then sorted into one of three groups; indels, exonic nsSNVs, and any variant within 25 base pairs of an exon. All other variants were filtered out.

Indels and exonic nsSNVs were analysed using Alamut Visual mutation analysis software (<http://www.interactive-biosoftware.com/alamut-visual/>). For nsSNVs this gave both PolyPhen2 and SIFT scores that predicted pathogenicity. Variants scored as benign by PolyPhen2 and/or tolerated by SIFT were filtered out. Indels were analysed for their potential frameshift effect. Those that caused a stop codon downstream were retained, while the rest were filtered out.

NGS technologies in general show less accuracy detecting indels than they do detecting nsSNVs. Indels are expected to be highly damaging due to their frameshift effect on sequence that creates stop codons and changes the translation of downstream codons, yet many are identified in any sequencing run, which seemingly have no effect.

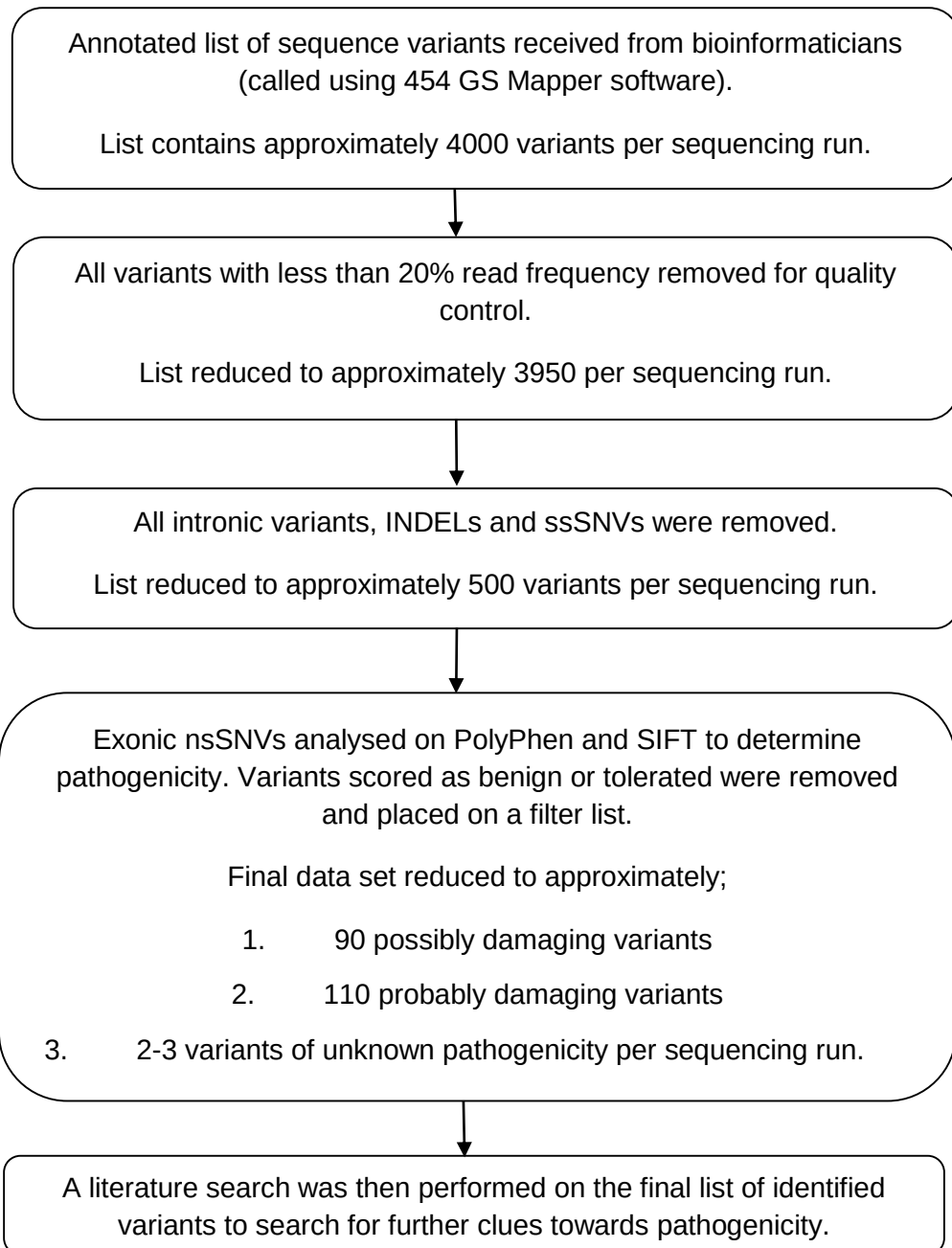


Figure 4.12 Initial data analysis pipeline. This was used for the first batch of samples sequenced.

It seems highly unlikely that all these mutations can be real, and often on analysis of the raw reads using the GS Reference Mapper software many turned out to be incorrect calls of SNVs in close proximity.

The filtering and analysis pipeline was also improved with the use of splice prediction software to analyse the effect of mutations that could potentially have an effect on the splicing of an intron. A number of programs have been developed that attempt to predict whether a mutation causes a change to the splicing of a genomic sequence (listed in **Table 2.1** p. 73). None of these programs are entirely accurate though, and most will give many false positives for the prediction scores of variants analysed with them.

Variants within 25bp of an intron/exon boundary were analysed on Alamut using four different splice prediction programs; NNSPLICE, Human Splice Finder, MaxEntScan, and SpliceSiteFinder-Like. Variants that gave less than a 10% change in the splice scores of at least three programs were filtered out. Variants that directly changed a GU splice donor or AG splice acceptor site were not filtered out and were considered to potentially have an effect on splicing.

All the remaining variants were annotated with their frequency in ethnically matched control populations using the HapMap database which contains data from sources such as the 1000 genomes project. Variants that appeared at more than 20% frequency in control populations were filtered out as the majority of the disease phenotypes investigated are rare, so it is unlikely that variants causing them would appear at such a high frequency in healthy populations.

Finally, as in the initial pipeline, a literature search was performed on the remaining variants to determine whether they had previously been identified as disease causing.

Where available a family history of the disease was taken and relatives tested to allow confirmation of cosegregation of the variants in the final list with the disease phenotype. In some cases, this analysis could preclude a variant from being considered, as it should not appear in healthy relatives, and similarly it should be present in every affected relative.

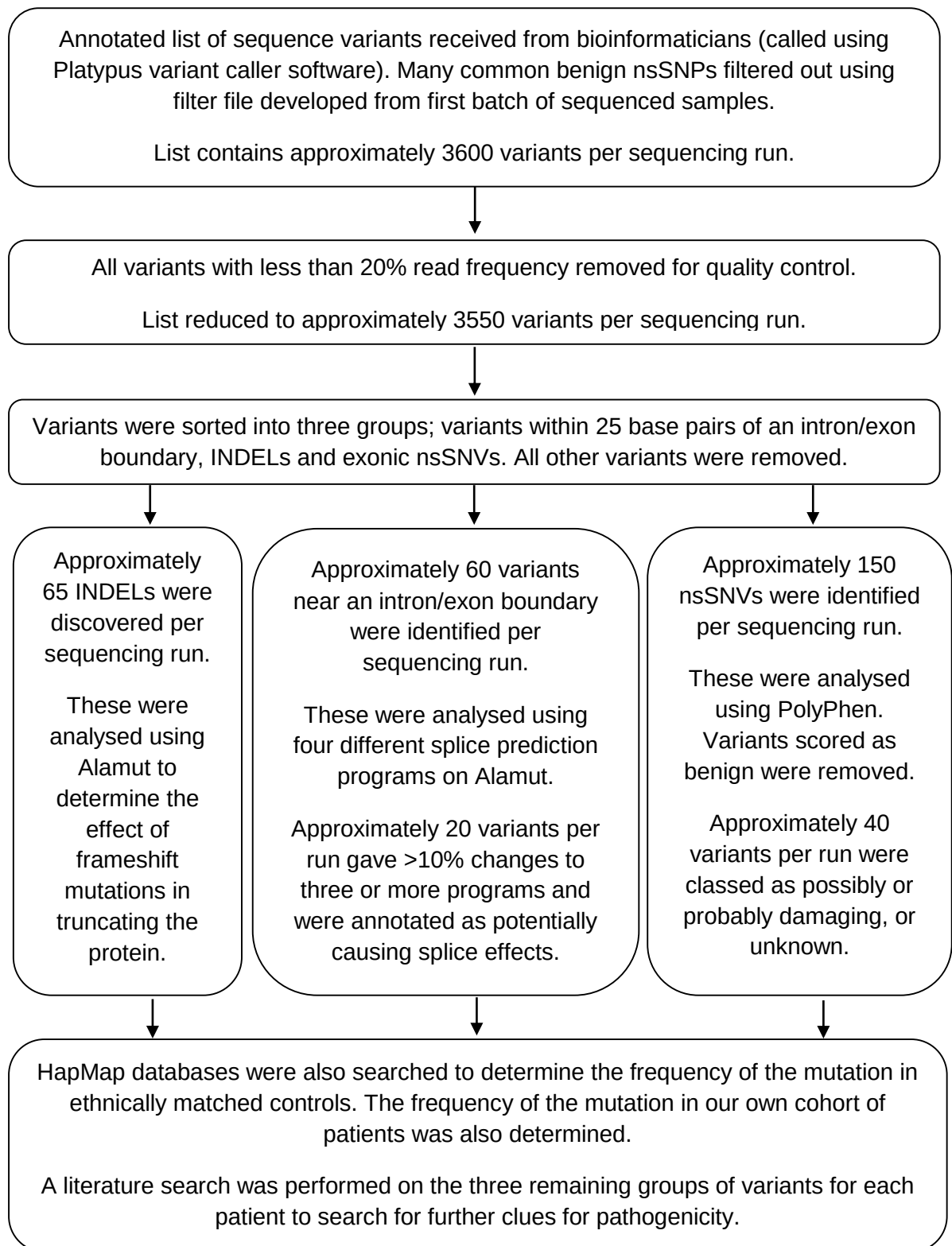


Figure 4.13 Improved data analysis pipeline. This was used for the second batch of samples sequenced. This pipeline included analysis of INDELs and potential splice site changes.

4.3.3 Roche 454 sequencing results

In total 55 patient samples with a range of IRD diseases were sequenced. Of these 55 samples, five failed to produce high enough quality data, leaving 50 samples to be analysed.

The first 28 samples sequenced were analysed using the initial bioinformatics pipeline shown in **Figure 4.12**, which did not include analysis of INDELs and Intronic variants. The results of this filtering are shown in **Table 4.1**.

Having analysed the results of this pipeline, the improved bioinformatics pipeline shown in **Figure 4.13** was developed, giving the final data shown in **Table 4.2**. Subsequently the initial 28 patients were also analysed for variants in INDELs, and those near splice sites, to determine whether variants had been missed in their analysis.

Table 4.1 Summarised sequencing results for first batch of 28 patients analysed using initial pipeline.

Mean number of exonic variants discovered per patient, sorted by PolyPhen2 score	Possibly damaging	7.5
	Probably damaging	9.1
	Unknown	0.2
Mean sequencing metrics data per patient	Proportion of reads in array	9.4%
	Alignment to ROI	57.4%
Mean coverage per patient	bases > 10 reads	92.6%
	bases > 30 reads	70.7%
	Mean number of reads	55.9

Table 4.2 Summarised sequencing results for next 22 patients analysed using second pipeline.

Mean number of variants discovered per patient	Possibly damaging exonic nsSNV	1.2
	Probably damaging exonic nsSNV	2.1
	Unknown exonic nsSNV	0.1
	INDELs	5.3
	Potential splice effect	0.2
Sequencing metrics data	Proportion of reads in array	9.1%
	Alignment to ROI	63.3%
Coverage	bases > 10 reads	95.9%
	bases > 30 reads	75.2%
	mean number of reads	49.2

4.3.4 Pathogenic variants identified

The 50 patients who were analysed included 36 “unknowns” diagnosed clinically with either retinitis pigmentosa (RP), or cone-rod dystrophy (CRD). The remaining 14 patients were positive controls in whom a secure molecular diagnosis had already been found using Sanger sequencing.

Of the 14 control patient samples, 12 were compound heterozygotes or homozygous for pathogenic mutations in *ABCA4*, while the other two had heterozygous dominant mutations in the *PRPF8* and *PRPF31* genes. All of these known variants were correctly identified by the next-generation sequencing and the bioinformatic pipeline developed for filtering.

Out of the patients who had received no previous genetic testing, the bioinformatics pipeline was able to identify a secure molecular diagnosis for nine (25%) of the patients. The pathogenic mutations identified in each of the nine patients are listed in **Table 4.3**. This data was published in the *European Journal of Human Genetics* (Shanks *et al.* 2013).

4.3.4.1 ATP-binding cassette transporter variants in patients 82, 86, and 92

Compound heterozygous *ABCA4* mutations were identified in patient 82 who had been diagnosed with RP at the age of 11. One chromosome contained a heterozygous mutation c.1906G>A (p.Gln636X) in exon 13, which does not appear in the ExAC database. This nonsense mutation causes premature termination of translation, resulting in the loss of approximately 70% of the protein. The second variant was a c.768C>A (p.Val256Val) heterozygous substitution in exon 16 which appears at 0.0001072 allele frequency in the ExAC database. The mutation is synonymous but is present on the last base of the exon and is predicted to affect splicing by skipping exon 16 (NNSPLICE 5' splice score changes from 1.0 to 0, GeneSplicer changes from 4.0 to 0). This mutation was only identified in the second data analysis pipeline, in which potential splice mutations were analysed. The mutation has previously been reported in several RP patients by September and colleagues (September *et*

al. 2004). The severity of the RP phenotype and the early age of onset would be expected from such severe mutations in the *ABCA4* gene.

A homozygous c.2894T>C (p.Asn965Ser) substitution was identified in exon 19 of the *ABCA4* gene in patient 86, which was predicted to be Probably damaging by PolyPhen2 (score 0.999), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value -1), with a 0.0001731 allele frequency in the ExAC database. This mutation was shown by Sun and colleagues to show “little or no retinal-stimulated ATP hydrolysis” (Sun *et al.* 2000). Patient 86 was diagnosed with an RP phenotype with an early age of onset, which would be expected from variants on both alleles of the *ABCA4* gene with such severe effect.

Patient 92 had a phenotype of severe CRD. A homozygous c.6658G>A (p.Gln2220X) substitution was identified in exon 48 of the *ABCA4* gene in this patient, which appears at 0.00004946 allele frequency in the ExAC database. This nonsense mutation creates a stop codon which causes premature termination of translation, resulting in a protein of 2200 amino acids in length instead of the 2273 found in the wild type. The mutation had previously been reported as pathogenic by Maugeri and colleagues (2000).

4.3.4.2 Guanylate cyclase 2D variant in patient 94

Patient 94 had been diagnosed with cone-rod dystrophy at 15 years of age. A single heterozygous c.2513G>A (p.Arg838His) substitution was identified in exon 13 of the *GUCY2D* gene of the patient, which was predicted to be Probably damaging by PolyPhen2 (score 1.000), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value -0.999). The mutation did not appear in the ExAC database. This variant was previously reported by Weigell-Weber and colleagues, who showed that this codon is present in the putative dimerization domain of the RET-GC1 protein (the previous name for *GUCY2D* was *RETGC1*), and as a result a mutation here could cause an inability to regenerate cyclic guanosine monophosphate (Weigell-Weber *et al.* 2000). Dominant mutations in the *GUCY2D* gene have previously been

reported to cause cone-rod dystrophy (Kelsell *et al.* 1998). The patient had an affected daughter but no previous family history in the parents, suggesting the mutation was *de novo* in patient 94.

4.3.4.3 Phosphodiesterase 6B variants in patients 95 and 123

Compound heterozygous mutations in the *PDE6B* gene were identified in patient 95 who had been diagnosed with RP. The first was a c.892C>T (p.Gln298X) nonsense substitution in exon 5 which causes premature termination of translation, resulting in the loss of approximately 87% of the protein. The variant appears at 0.00004157 allele frequency in the ExAC database.

The second mutation was a c.2193+1G>A in intron 18, which removes the guanine base from the GU splice donor site, which is predicted to cause skipping of exon 18, and resulted in the complete removal of the 5' splice donor prediction score for all five of the splice prediction programs used. This variant does not appear in the ExAC database.

Compound heterozygous mutations in the *PDE6B* gene were also identified in patient 123 who also had RP. The first was a c.163G>T (p.Glu55X) substitution in exon 1, which does not appear in the ExAC database. This nonsense mutation creates a stop codon at amino acid 55 which causes premature termination of translation, resulting in the loss of approximately 90% of the protein. The second pathogenic mutation is the same c.2193+1G>A splicing mutation in intron 18 which was identified in patient 95.

The *PDE6B* gene codes for rod phosphodiesterase which is involved in the hydrolysis of cyclic guanosine monophosphate (cGMP) in the phototransduction cascade (Khrantsov *et al.* 1993). Variants in *PDE6B* have been reported by McLaughlin and colleagues (1993) to be the cause of RP.

4.3.4.4 *Crumbs family member 1 variants in patient 100*

Compound heterozygous mutations in the *CRB1* gene were identified in patient 100 who had been diagnosed with Leber congenital amaurosis (LCA). One of the variants identified was a substitution at c.2234C>A (p.Thr745Lys) which was predicted to be Probably damaging by PolyPhen2 (score 1.000), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value -1). The variant had a 0.00008263 allele frequency in the ExAC database. The mutation had not previously been reported in the literature, however another mutation in this codon c.2234C>T (p.Thr745Met) had been previously reported by den Hollander and colleagues, and it was considered pathogenic due to the conservation of this amino acid (den Hollander *et al.* 1999). The same group went on to identify the c.2234C>T (p.Thr745Met) mutation repeatedly in a larger cohort of patients with a spectrum of inherited retinal dystrophies, including both Leber congenital amaurosis and RP (den Hollander *et al.* 2004).

The second pathogenic mutation identified in the *CRB1* gene of patient 100 was a c.2842+5G>A substitution in intron 8 which was predicted to affect splicing (three of the five prediction program scores for the nearby 5' donor site disappeared completely with the change), causing skipping of exon 8.

The *CRB1* gene codes for the CRB1 protein, which is homologous to the *Drosophila* transmembrane crumbs protein (den Hollander *et al.* 1999). The mechanism by which mutations in the *CRB1* gene cause retinal dystrophies is not fully understood, however epithelial cells lacking the gene fail to organise the zonular adherens which binds cells together, and allows them to maintain orientation and polarity (Tepass and Knust 1993, Grawe *et al.* 1996).

4.3.4.5 *Rhodopsin variant in patient 111*

In patient 111 a heterozygous c.158C>G (p.Pro53Arg) substitution was identified in exon 1 of the rhodopsin gene (*RHO*) which was predicted to be Probably damaging by PolyPhen2 (score 1.000), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value -1).

The variant did not appear in the ExAC database, but had been reported in a publication by Inglehearn and colleagues (1992). The mutant protein produced by this variant has been classed as clearly defective, due to the differing results it produced on three different assays compared to the wild type protein, and the retention of the mutant protein in the endoplasmic reticulum (Sung *et al.* 1993). The mutation in patient 111 was *de novo*, making this one of the first *de novo* variants to be reported in *RHO*.

The phenotype of patient 111 had previously been considered to be a recessive RP based on the age of onset at three years of age, and the lack of family history. The diagnosis of a dominant mutation changed the genetic counselling, as the patient's children are at 50% risk from inheriting the condition, in contrast to the relatively low risk of inheriting a recessive condition.

4.3.4.6 Cone-rod homeobox variant in patient 121

Patient 121 had been diagnosed with macular dystrophy at the age of 17. A c.774T>G (p.Tyr258X) substitution was identified in exon 4 of the cone-rod homeobox (*CRX*) gene, which does not appear in the ExAC database. This nonsense mutation causes the loss of 30% of the protein including the OTX domain at the carboxy terminus. Loss of this domain has been shown by Chau and colleagues to cause pathogenicity by preventing the mutant protein from initiating the rhodopsin promoter (Chau *et al.* 2000). Familial segregation testing confirmed that the variant was present in all three affected siblings of the patient.

Table 4.3 Pathogenic mutations identified through the analysis pipeline.

Patient ID	Diagnosis	Age of onset	1st mutation		2nd mutation		Evidence of pathogenicity	Family history	Co-segregation
			Gene	Variant	Gene	Variant			
82	RP	11	<i>ABCA4</i>	c.1906G>A (p.Gln636X)	<i>ABCA4</i>	c.768C>A (p.Val256Val)	E636X is a premature termination codon that causes the loss of almost 70% of the protein. V256V results in a predicted splice site mutation.	Simplex	-
86	RP	10	<i>ABCA4</i>	c.2894T>C (p.Asn965Ser)	<i>ABCA4</i>	c.2894T>C (p.Asn965Ser)	Homozygous mutations result in a protein that shows very little retinal-stimulated ATP hydrolysis.	Affected sister and consanguineous first cousin parents	Sister homozygous
92	Cone-rod dystrophy	8	<i>ABCA4</i>	c.6658G>A (p.Gln2220X)	<i>ABCA4</i>	c.6658G>A (p.Gln2220X)	Homozygous premature termination codon mutations that causes the loss of 53 downstream AA.	Two affected male cousins and consanguineous first	Parents heterozygous
94	Cone-rod dystrophy	15	<i>GUCY2D</i>	c.2513G>A (p.Arg838His)	N/A (dominant)		Mutant less sensitive to calcium concentrations, causing excess calcium to build up in photoreceptors causing apoptosis.	Affected daughter	-
95	RP	18	<i>PDE6B</i>	c.892C>T (p.Gln298X)	<i>PDE6B</i>	Intron 18 c.2193+1 G>A	Q298X is a premature stop codon causing loss of 35% of the protein. The intron 18 c.2193+1 G>A variant removes the G from a GU splice donor site.	Simplex	-
100	LCA	4 months	<i>CRB1</i>	c.2234C>A (p.Thr745Lys)	<i>CRB1</i>	Intron 11 c.2842+5 G>A	T745M was reported to disrupt an area of cytoplasmic domain that was found to be crucial to function in <i>Drosophila</i> . The intron 11 c.2842+5 G>A variant is predicted to affect splicing.	Simplex	-
111	RP	3	<i>RHO</i>	c.158C>G (p.Pro53Arg)	N/A (dominant)		P53R expressed at low levels, regenerated chromophore poorly and was retained in the ER.	Simplex	<i>de novo</i> mutation
121	Macular dystrophy	17	<i>CRX</i>	c.774T>G (p.Tyr258X)	N/A (dominant)		Premature termination codon results in the loss of the OTX domain at the carboxy terminus. Chau <i>et al.</i> showed this loss prevents the mutant protein being able to initiate rhodopsin promoter.	Three affected siblings	Present in all three affected
123	RP	14	<i>PDE6B</i>	c.163G>T (p.Glu55X)	<i>PDE6B</i>	Intron 18 c.2193+1 G>A	E55X is a premature termination codon that causes the loss of almost 90% of the AA, including the cGMP binding domain and the catalytic domain. The intron 18 c.2193+1 G>A variant removes the G from a GU splice donor site.	Simplex	-

Key: RP = Retinitis pigmentosa, LCA = Leber congenital amaurosis

4.3.5 Exome sequencing analysis of patient 525A6

Patient 525A6 was originally diagnosed with Stargardt disease but much later clinical review suggested an end-stage RP of unknown molecular cause. The patient also had congenital myasthenic syndrome, a condition which has never previously been reported in conjunction with an RP phenotype. The myasthenic syndrome was found by another research lab to be caused by compound heterozygous mutations in the *GFPT1* gene (Senderek *et al.* 2011).

To investigate the cause of the early onset RP, the *ABCA4* gene was sequenced in this patient in the ORGL, but no pathogenic variants were identified.

The patient then underwent targeted sequencing using the NimbleGen sequence capture array and Roche 454 sequencing for 74 genes associated with retinal disease, and a single c.133T>C (Phe45Leu) *RHO* mutation was identified which had a 0.000008237 allele frequency in ExAC. The effect of this mutation was predicted to be Probably damaging by PolyPhen2 (score 0.988), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value 1).

However, functional testing of this variant showed that the activity of the protein produced was not significantly different to the wild type, and so the variant was unlikely to be the cause of the retinitis pigmentosa (Davies *et al.* 2012). This was supported by a later study by another group who reported the variant as not causing the condition in a family with autosomal dominant retinitis pigmentosa (Vincent *et al.* 2013). This variant is also unlikely to be pathogenic as it is present in the unaffected father.

Array CGH was also performed by the ORGL and was shown to be negative (in case the myasthenic syndrome and RP were caused by a single deletion or other abnormality).

Owing to the negative results from the tests so far performed, it was decided that this patient should be tested using exome sequencing. The complex phenotype would provide an advantage in the analysis, as the known *GFPT1* variants could act as positive controls to test the bioinformatic pipelines used. The sequencing was performed on the patient and both healthy parents by the High Throughput Sequencing group at the Wellcome Trust Centre for Human

Genetics, using the Illumina HiSeq 2000 with a SureSelect whole exome capture kit, and a list of variants was generated which included the phase of the variant according to parental origin.

The far greater number of variants that were identified in the patient by exome sequencing compared to variants identified in the 74 genes analysed in the previous parts of this chapter required the use of a different analysis pipeline, so this was used as an opportunity to test three newly released bioinformatic programs for variant pathogenicity prediction. The programs used were:

1. PolyPhen2 batch upload (in combination with SIFT)
2. Alamut batch
3. Ingenuity Variant Analysis

4.3.5.1 PolyPhen2 batch upload

The variant called file generated by exome sequencing was uploaded to PolyPhen2 using the batch upload feature (<http://genetics.bwh.harvard.edu/pph2/bgi.shtml>). The results file produced was filtered to remove all variants in genes that were not in the 194 genes listed on RetNet (an online database of retinal genes) at that time as having an association with retinal dystrophies. This filtering left 192 nsSNVs, of which:

- 138 were scored as benign
- 20 were scored as possibly damaging
- 28 were scored as probably damaging

These variants were then also analysed on SIFT (<http://sift.jcvi.org/>). The only variant that was scored as either possibly or probably damaging by PolyPhen2, scored as deleterious by SIFT, and which was also in a retinal gene, was the c.133T>C (Phe45Leu) *RHO* variant previously found using the targeted analysis.

This analysis pipeline was repeated for genes involved in congenital myasthenic syndrome and was able to correctly identify the *GFPT1* mutations responsible for this condition.

4.3.5.2 Alamut Batch

The entire Phased data file was also uploaded to Alamut Batch (available at <http://www.interactive-biosoftware.com/alamut-batch/>). This program automates the annotation of a variant called file (VCF), with all the information the original Alamut Visual software provides for individual variants. The resulting annotated VCF generated was then filtered to remove any variants not present in the 194 genes listed on RetNet at that time as having an association with retinal dystrophies. This gave 436 variants, six of which were annotated as “pathogenic” and nine as “probable-pathogenic” in the rsClinicalSignificance data field by the software. Of these fifteen, only one was also identified as being deleterious on SIFT, the c.133T>C (Phe45Leu) variant in *RHO*. As was done with PolyPhen2 batch upload, this analysis pipeline was repeated for genes involved in congenital myasthenic syndrome and it was found to be able to correctly identify the *GFPT1* mutations responsible for this condition.

4.3.5.3 Ingenuity Variant Analysis

The entire phased variant called file was uploaded to Ingenuity Variant Analysis software, which automatically filters VCFs to identify pathogenic mutations based on disease phenotype keywords inputted by the user. The data was analysed with the keywords congenital myasthenic syndrome, which correctly identified the heterozygous mutations in *GFPT1* (**Figure 4.14**).

When this analysis was repeated for the keyword “retinitis pigmentosa”, “RP”, and for a variety of other related disease phenotype keywords, a list of 7 heterozygous variants were identified by the software (**Figure 4.15**). For each of these seven variants at least one of the unaffected parents had the same mutation (shown by the boxes in the “control samples” column), with no evidence of a second compound heterozygous variant in each of the seven genes. This suggested that these seven variants did not cosegregate with the disease, and so are unlikely to be the cause of the RP phenotype. It is not known why these variants were left in the final list, as they should have been removed by the genetic analysis filter. This final list again included the c.133T>C (Phe45Leu) *RHO* variant identified by the previous two analysis pipelines.

4.4 Discussion

In the work described in this chapter the NimbleGen sequence capture array was combined with the Roche 454 sequencing platform, to sequence and analyse DNA samples from a cohort of 36 patients with inherited retinal dystrophies (IRD) in whom no previous genetic diagnosis had been made. A bioinformatic pipeline was developed to filter and analyse the data produced from the sequencing.

The pipeline developed utilised *in silico* methods to predict the pathogenicity of variants, as well as a literature search to determine whether variants had been previously reported. The pipeline was able to identify a secure molecular diagnosis in 9 out of 36 (25%) of patients using these methods (**Table 4.3**), showing that NGS is an effective tool for molecular diagnosis of genetic diseases. However, the low rate of 25% illustrates that the methodology and technology required significant improvements.

In the remaining 27 patients there are a number of possible reasons why the pathogenic variants were not identified. Firstly, the capture design used for the targeted sequencing of these patients included just 74 genes which were known at the time to have an association with retinal dystrophies. This number was chosen due to technical limitations of the capture technologies used, which prevented more genes from being targeted. The same design was used for each of the technologies to allow a comparison of what was identified by each. In the time since that design was made a further 100 genes have been identified as having an association with retinal disease, as shown in **Figure 1.27** in section 1.8.2 p. 65.

Increasing the number of genes in the capture design significantly increases the molecular diagnosis rate. For example one report using SureSelect capture of 254 retinal disease genes with sequencing on the Illumina Genome Analyzer IIX gave a higher detection rate of 57% (Audo *et al.* 2012).

The discovery of new genes and gene associations, and the difficulty of re-designing targets for sequence capture technologies, is one of the biggest disadvantages of using targeted

sequencing for molecular diagnostics, and is one of the main reasons why this type of sequencing has since been supplanted by exome and whole genome sequencing.

Other potential reasons for the failure to detect causal variants are poor coverage of the target region, or another technical flaw in the sequencing chemistry or detection methods prevented detection. A variant may also be mosaic: present only in the tissue of the retina, and not the blood samples from which DNA was extracted, which could make it either impossible to detect from such samples, or extremely difficult to detect at low levels of coverage. If filtering is performed at less than 20% of the variants in order to identify potential mosaic cases, then the numbers of false positives is likely to increase.

Finally, it may be that the variants have been sequenced and identified, but the analysis pipeline used was not able to recognise them as pathogenic, leading to them being filtered out. This is more likely to be the case for variants that have not previously been reported in the literature, or for variants that are located in genes not currently known to have an association with a specific clinical phenotype, as in either case they would not be found in a literature search.

Neveling and colleagues used a similar methodology of sequencing to that reported in this project (NimbleGen sequence capture array with sequencing on the 454 GS FLX) and captured 111 genes, reporting a higher estimated detection rate of 50% for screening RP patients with no genetic diagnosis (Neveling *et al.* 2012).

In addition to capturing more genes another possible reason for the discrepancy in detection rate between their results and those reported in this chapter is the method of determining pathogenicity. In their data analysis pipeline three pathogenicity prediction programs were used for missense variants, with a simple majority verdict used to determine pathogenicity (Neveling *et al.* 2012). This would have led to a greater number of false positives compared to our pipeline, in which missense variants were only considered pathogenic with a positive result on both prediction programs used, and the presence of supporting evidence in the literature.

Conversely, our initial assessments of pathogenicity may have been too stringent, especially when considering novel missense variants which may have been discounted because of lack of functional or bioinformatic evidence of pathogenicity.

There have been some reports made that claim to have high detection rates, however the claimed figures refer to the number of pathogenic variants detected, and not the number of patients in whom a secure molecular diagnosis was found. For example one report on diagnosis of inherited retinal disease claimed a 50% detection rate, however this figure included patients with only a single recessive pathogenic mutation identified (O'Sullivan *et al.* 2012).

In the case of patient 525A6 the patient also underwent exome sequencing after the targeted sequencing failed to reveal a molecular diagnosis. Sequencing the exome should in theory identify the variants in all known genes in the genome of a patient. However, several analyses of the exome sequencing data failed to identify a potentially pathogenic variant in a known retinal disease gene. This might be because the variant was either called by the sequencing data, but not identified as pathogenic by the analysis pipeline, or it could be due to the variant being present in a gene which although sequenced, has not yet been identified with retinal disease, and so was also excluded by the filtering steps. Finally the variant could have been present in a gene which was not targeted by the exome capture design, as these designs are known to vary depending on the manufacturer, with each one capturing sequence that others do not (Sulonen *et al.* 2011).

A further discussion of why pathogenic variants are not always identified by next-generation sequencing is presented in the introduction to Chapter 6 (section 6.1.4 p.209).

4.4.1 Conclusions

In this chapter it was shown that NGS technologies can identify pathogenic variants in patients with genetic diseases, however there are great challenges in analysing the data generated, and developing bioinformatic pipelines that are capable of filtering out benign variants to identify the very few number of variants that are actually pathogenic.

Many *in silico* programs have been developed such as PolyPhen2, SIFT, Alamut, and the splice predictions programs used in this chapter. Newer programs have been developed that allow mass upload of NGS data such as PolyPhen 2 batch upload, and High Throughput Alamut. More recently some programs, such as Ingenuity Variant Analysis, aim to automate the entire process, allowing a researcher to upload a raw VCF to a server, specify a keyword for the disease phenotype, and then have the filtering performed for them automatically.

These programs were found to be useful and in some cases could quickly identify pathogenic variants, however no *in silico* method yet developed has a 100% detection rate. Programs that rely on *in silico* prediction of the effect of a mutation on protein structure had high rates of false positive calls. They were also usually designed to be used just on single nucleotide substitutions, rather than deletions, insertions or rearrangements.

Programs that rely on curated databases of known pathogenic mutations, such as Ingenuity Variant Analysis, will not be able to identify a mutation if it has not been previously reported, as was shown in the analysis of patient 525A6 where the variants responsible for the myasthenic syndrome phenotype were identified, but the variants responsible for the RP phenotype were not.

The next chapter describes experiments performed to provide further evidence to either support or disprove the pathogenicity of variants identified by NGS. This can be done by testing for cosegregation of the variant with the condition in family members, and in some cases by testing mRNA to determine the effect a genomic variant might have on transcription.

CHAPTER 5

Testing variants identified by whole genome sequencing in patients with early onset epilepsy

5.1 Introduction

The previous two chapters investigated the ability of next-generation sequencing to detect pathogenic variants in retinal dystrophies. One of the challenges presented by the analysis of the data in Chapter 4 was the need to filter out the many benign variants identified from the few pathogenic variants of interest. Although bioinformatic programs for determining pathogenicity were able to provide initial filtering of the raw data, there were still several benign variants remaining for each patient after filtering with these methods. Further analysis to prove pathogenicity may require more direct functional experimental data to determine which of the candidates the causal pathogenic mutation is.

An opportunity arose to perform such experiments in another project that was occurring concurrently with the retinal project, which was using whole genome sequencing (WGS) to investigate epilepsy syndromes. Performing further analysis is especially important for exome or WGS data, as the number of candidate variants identified increases greatly.

The investigations were performed on two patients (OTH 12 and OTH 18). WGS had been performed by the WGS 500 group at the Wellcome Trust Centre for Human Genetics (University of Oxford) on samples from both these patients and their parents, and an analysis of these data was performed using Ingenuity Variant Analysis software by Hilary Martin (University of Oxford). Four potentially pathogenic variants in genes known to be associated with early onset epilepsy were identified; three variants in patient OTH 12, and one variant in patient OTH 18.

The work described in this chapter consisted of firstly confirming the presence of the variants in the patients using Sanger sequencing, and that they cosegregated with the disease in family members. Secondly, experiments were performed on two of the genomic variants to test whether they affected splicing. This involved extracting mRNA from leukocytes, reverse transcribing it into cDNA, amplifying regions containing the variants, and sequencing the products across the exons that were predicted to be skipped due to the variants detected.

5.1.1 Early onset epilepsy

Epilepsy refers to a condition in which a patient suffers from recurrent unprovoked seizures, which are caused by an imbalance in excitation and inhibition of neuronal polarisation (Matsumoto and Marsan 1964, Prince and Wilder 1967, Walther *et al.* 1986), and as such the majority of genes so far discovered that cause epilepsy affect control of neuronal polarisation (Staley 2015). A number of different forms of epilepsy that commence in the first year of life have been identified, and several genes have been reported to have an association with these conditions.

Ohtahara syndrome, also known as Early Infantile Epileptic Encephalopathy with suppression burst (EIEE, OMIM #308350) is an epileptic encephalopathy that affects newborns (Ohtahara *et al.* 1997). Clinically it is characterised by frequent tonic spasms, and a burst suppression pattern on an electroencephalogram (EEG) in which periods of high voltage bursts of activity alternate with periods of nearly flat suppression phases, as shown in **Figure 5.1**. The disorder is debilitating, causing severe mental retardation, and is often fatal.

Approximately 75% of patients with Ohtahara syndrome progress to West Syndrome, a condition characterised by infantile spasms, an EEG abnormality known as hypsarrhythmia and psychomotor delay or arrest (West 1841, Wheless *et al.* 2012).

In 2004 Kato and colleagues showed that Ohtahara syndrome can be caused by mutations in the aristaless-related homeobox (*ARX*) gene, which is also responsible for a spectrum of other epilepsies and neurological disorders (Kato *et al.* 2004). They later reported that the disease can be caused by a specific alanine repeat expansion in the *ARX* gene (Kato *et al.* 2007).

Another type of early onset epilepsies are benign familial neonatal seizures (BFNS, OMIM #121200), which occur within the first few days of life, but then generally recede by one year of age (Rett and Teubel 1964). BFNS have been shown to be caused by mutations in *KCNQ2* and *KCNQ3*, genes that encode the voltage-gated potassium channel subunits K_v7.2 and K_v7.3 respectively (Biervert *et al.* 1998, Charlier *et al.* 1998, Singh *et al.* 1998).



Figure 5.1 Electroencephalogram (EEG) trace showing a burst suppression pattern consistent with Ohtahara syndrome. A burst suppression is characterised by periods of high voltage neurological activity in the brain (burst phase) alternating with periods of no activity (suppression phase).

Benign familial neonatal-infantile seizures (BFIS3, OMIM #607745) can develop later than BFNS, within the first four months, and have been shown to be caused by mutations in *SCN2A* which encodes the $\alpha 2$ -subunit of the voltage-gated sodium channel (Heron *et al.* 2002).

Benign familial infantile seizures (BFIS) start later still between four and eight months of age, and have been mapped to two different loci on chromosomes 16 and 19 (Guipponi *et al.* 1997, Caraballo *et al.* 2001), but the causative genes are yet to be identified.

A range of epileptic syndromes with febrile seizures (convulsions with raised body temperature) have also been identified, of which Ohtahara syndrome is one. Generalised epilepsy with febrile seizures plus, type 1 is caused by mutations in the *SCN1B* gene (Wallace *et al.* 1998), type 2 by mutations in the *SCN1A* gene (Escayg *et al.* 2000), and type 3 by mutations in the *GABRG2* gene (Baulac *et al.* 2001).

The extensive heterogeneity of epileptic conditions in infants and neonates makes these conditions suitable candidates for use with next-generation sequencing (NGS) technologies enabling a large number of genes to be screened.

The work described in this chapter focussed on two patients who will be referred to as “OTH 12” and “OTH 18”. Patient OTH 12 had Ohtahara syndrome and severe developmental delay. Patient OTH 18 had Severe Non-Syndromic Early Onset Epilepsy, severe developmental delay, patent ductus arteriosus, and atrial septal defect as a neonate, as well as cutaneous hypopigmentation. Clinical phenotypes were diagnosed by Zenobia Zaiwalla in the Department of Clinical Neurophysiology, and Tony McShane in the Department of Paediatrics, both based at the John Radcliffe Hospital (University of Oxford).

Sanger sequencing of the *ARX* gene and several other known Ohtahara genes had not identified a causal mutation in these patients, so whole genome sequencing was performed on both of these patients and their parents by the high-throughput sequencing group at the Wellcome Trust Centre for Human Genetics. Four potentially pathogenic variants were identified by Hillary Martin using Ingenuity Variant Analysis software.

5.1.2 Preliminary analysis of identified variants in patients OTH 12 and OTH 18

This chapter describes experiments performed to provide further evidence that the variants detected in these two patients with early onset epilepsy are pathogenic. The aims of the work were to first confirm the variants identified by WGS were present in the patients, and to test cosegregation in family members using Sanger sequencing. Variants that were confirmed by Sanger sequencing and which showed segregation with other family members in these results were then tested further using mRNA extracted from the patients to determine whether exon skipping was being caused by a variant predicted to disrupt a splice site.

5.1.2.1 Variants identified in patient OTH 12 by whole genome sequencing

In patient OTH 12 three potentially pathogenic variants were identified by whole genome sequencing by Hilary Martin and the WGS 500 group:

1. A *de novo* heterozygous c.4367G>A (p.Ser1456Asn) substitution in the *C5orf42* gene predicted to be Probably damaging by PolyPhen2 (score 1.000), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value 0.953). The variant had an allele frequency of 0.000008237 in the ExAC database.

Mutations in this gene had recently been reported to cause Joubert syndrome (Srour *et al.* 2012), a disorder characterised by underdevelopment of the cerebellar vermis, and a “molar tooth sign” on an MRI indicating a malformed brain stem, often in conjunction with ataxia, hypernea, sleep apnea and hypertonia (Brancati *et al.* 2010). The clinician who diagnosed patient OTH 12 described phenotypes that show significant overlap with Joubert syndrome.

2. Compound heterozygous variants in the sodium channel, voltage-gated, type IV, alpha subunit (*SCN4A*) gene

- A c.2824G>A (p.Asp942Asn) which was predicted to be Probably damaging by PolyPhen2 (score 0.994), Tolerated by SIFT (score: 0.05) and Disease causing by MutationTaster (p-value 0.814) and with an allele frequency of 0.00005936 in ExAC.
- A c.5424G>A (p.Met1808Ile) predicted to be Benign by PolyPhen2 (score 0.000), Tolerated by SIFT (score: 0.27) and as a Polymorphism by MutationTaster (p-value 1) and with an allele frequency of 0.0001675 in ExAC.

A mutation in this gene had previously been reported to cause early onset epilepsy by changing inactivation of potassium channels (Tsujino *et al.* 2003).

3. A homozygous c.690-2A>G substitution in the essential splice acceptor site of intron 3 of the phosphatidylinositol glycan anchor biosynthesis, Class Q (*PIGQ*) gene which is predicted by all five splice prediction programs on Alamut to cause skipping of exon 3. *PIGQ* is part of the glycosphosphatidylinositol (GPI) biosynthesis pathway. Recessive mutations in several other genes involved in this pathway (*PIGA*, *PIGN*, *PIGM*, and *PIGV*) have been shown to cause epilepsy with seizures (Almeida *et al.* 2006, Krawitz *et al.* 2010, Maydan *et al.* 2011, Johnston *et al.* 2012).

The parents of patient OTH 12 were designated OTH 13 and OTH 14. DNA was also available from an unaffected sibling designated OTH 49 (**Figure 5.2 A**).

5.1.2.2 Variant identified in patient OTH 18 by whole genome sequencing

In patient OTH 18 only one change, a *de novo* c.1228-1G>A substitution in the essential splice acceptor site of intron 8 was identified by whole genome sequencing by Hilary Martin and the WGS 500 group in the proto-oncogene *CBL* at position. *CBL* is a regulator of the RAF/MAPK pathway (Swaminathan and Tsygankov 2006). Germline mutations in genes involved in this pathway can cause a range of developmental disorders known as cardiofaciocutaneous syndromes (Denayer and Legius 2007), disorders characterised by mental retardation and multiple congenital abnormalities which can include:

“failure to thrive, relative macrocephaly, a distinctive face with prominent forehead, bitemporal constriction, absence of eyebrows, hypertelorism, downward-slanting palpebral fissures often with epicanthic folds, depressed nasal root and a bulbous tip of the nose. The cutaneous involvement consists of dry, hyperkeratotic, scaly skin, sparse and curly hair, and cavernous haemangiomas. Most patients have a congenital heart defect, most commonly pulmonic stenosis and hypertrophic cardiomyopathy. The developmental delay usually is moderate to severe.” (Roberts et al. 2006).

The phenotype of these syndromes match the phenotypic diversity of the symptoms seen in patient OTH 18. The parents of patient OTH 18 were also available for testing, and were designated as OTH 19 and OTH 20 (**Figure 5.2 B**).

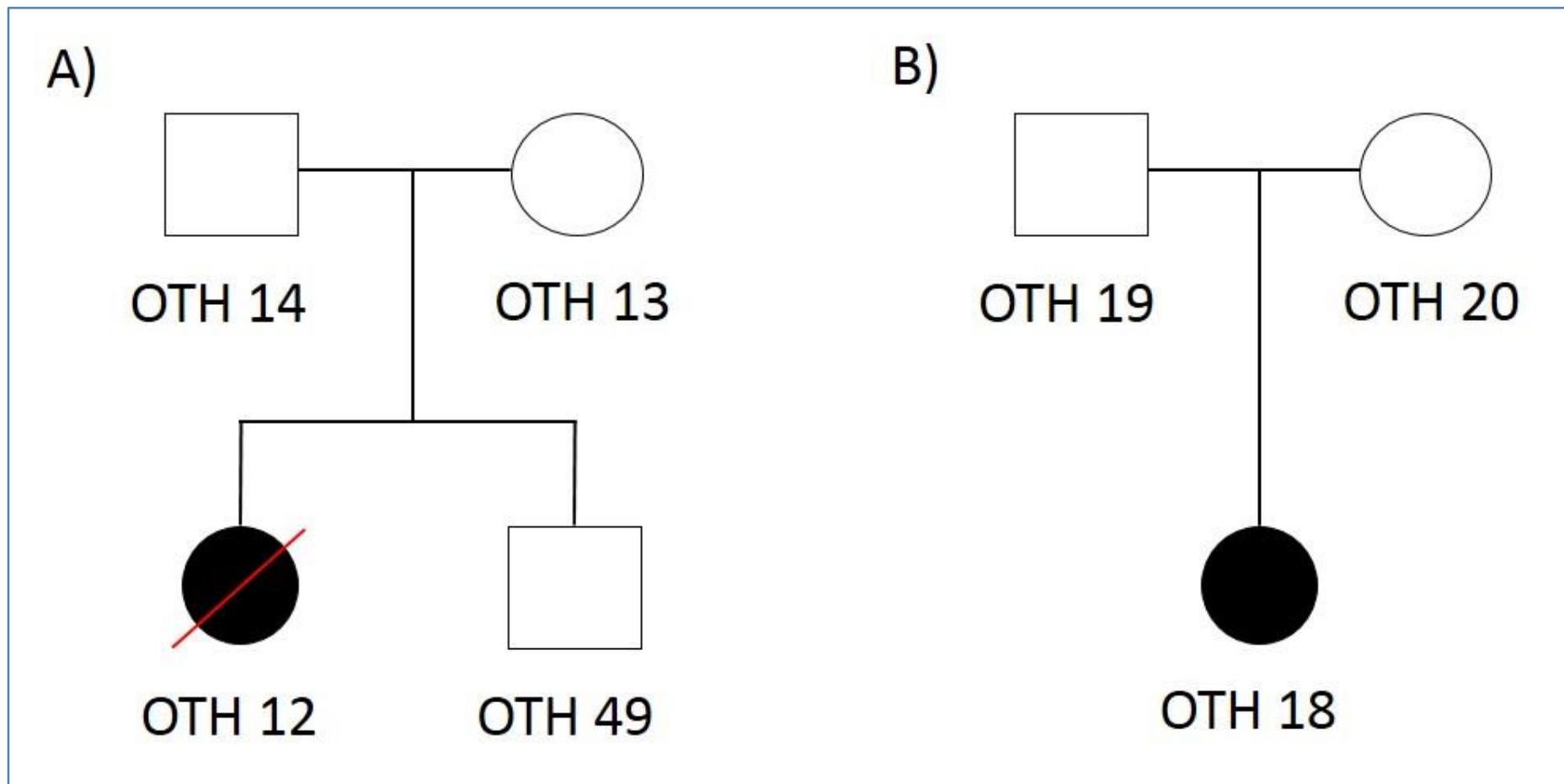


Figure 5.2 Pedigrees of the early onset epilepsy families. A) Patient OTH 12 had died from the condition. Both parents and a sibling were unaffected. B) Patient OTH 18 was affected and had no siblings.

5.2 Materials and Methods

DNA was extracted from whole blood samples of patients taken by venepuncture by staff at the John Radcliffe Hospital. DNA extraction was performed by the Oxford Regional Genetics Laboratory (ORGL) using Qiagen Autopure LS and EZ1 systems supplied by Qiagen (UK).

DNA samples were used as a template for amplification by PCR, and for Sanger sequencing, as described in sections 2.2.6 p. 79 and 2.2.8.1 p. 87 respectively.

5.2.1 mRNA extraction from blood

Paxgene Blood RNA Kits were supplied by Qiagen and contained PAXgene blood tubes, and Buffer BR1 (a resuspension buffer). Also supplied are Buffer BR2 (a binding buffer), and Buffer BR3 (a wash buffer), both of which contain guanidine isothiocyanate. Buffer BR4 is a wash buffer, which was supplied as a concentrate and required four volumes of 100% ethanol to be added before use. Buffer BR5 is an elution buffer. RNase-free water, and Proteinase K are also supplied in the kit. The process relies on column purification steps that bind nucleic acids and allow contaminants to be washed away.

Blood for mRNA extraction was taken from patients by venepuncture directly into a PAXgene tube (supplied by Qiagen UK) by clinical staff at the John Radcliffe Hospital.

The PAXgene tube containing blood was incubated at room temperature for between 12 and 24 hours. The PAXgene tubes contain a chaotropic buffer that destroys protein structures, and prevents mRNA degradation or transcription. The tube was then centrifuged for 10 minutes at 5000 x *g*, and the supernatant poured off and discarded. 5ml of RNase-free water was added and the tube re-sealed with a fresh hemogard. The tube was vortexed thoroughly to resuspend the pellet, then centrifuged for a further 10 minutes at 5000 x *g*. The supernatant was again poured off and discarded. The pellet was resuspended in 360µl of buffer BR1 by vortexing. The suspension was transferred to a 1.5ml tube and 300µl of BR2 and 40µl of Proteinase K were added. The tube was vortexed, and then heated at 55°C in a shaker-incubator at max speed for 10 minutes.

The tube was centrifuged at maximum speed for three minutes, the supernatant transferred to a fresh 1.5ml tube, and the pellet discarded. 350µl of 100% ethanol was added, and the tube was vortexed, then briefly centrifuged to bring the contents to the bottom. 700µl of the sample was added to a PAXgene column in a microcentrifuge tube and centrifuged at 8000 x *g* for one minute. The PAXgene column was transferred to a fresh microcentrifuge tube. The remainder of the sample was added to the PAXgene column, and then centrifuged at 8000 x *g* for one minute. The column was transferred to a fresh microcentrifuge tube.

700µl of buffer BR3 was added to the column which was then centrifuged for one minute at 8000 x *g*. The PAXgene column was placed in a fresh microcentrifuge tube, 500µl buffer BR4 was added, and the tube was centrifuged at 8000 x *g* for one minute. The supernatant was removed and discarded, and another 500µl of buffer BR4 was added, the tube spun at 8000 x *g* for three minutes, and the PAXgene column transferred to a fresh 1.5ml microcentrifuge tube.

The RNA was eluted by adding 40µl of buffer BR5 onto the column membrane, and centrifuging at 8000 x *g* for one minute. This elution step was repeated with another 40µl of buffer BR5. The eluted RNA was heated at 65°C for five minutes, and then chilled on ice. RNA quality was determined by nanodrop spectrophotometry (as described in section 2.2.3.1 p. 77). RNA was stored at -20°C until needed.

5.2.2 Reverse transcription and amplification of cDNA

Reverse transcription of mRNA into cDNA was performed using a QuantiTect reverse transcription kit from Qiagen (UK). Genomic DNA was removed by mixing 2µl of gDNA wipeout buffer with up to 1µg of RNA, made up to a total volume of 14µl with RNase free water and incubating the mixture for 2 minutes at 42°C. The reverse transcription reaction was made by mixing 1µl of Quantiscript Reverse Transcriptase with 4µl of Quantiscript RT Buffer 5x, and 1µl of RT primer mix. This mixture was then incubated at 42°C for 15 minutes, followed by inactivating the enzyme by heating at 95°C for 3 minutes, and allowing the reverse transcribed

cDNA to cool on ice. The cDNA generated was analysed using an Agilent 2100 Bioanalyzer with DNA 7500 chip (section 2.2.5 p. 78).

The cDNA was also used as a template for amplification by PCR and for Sanger sequencing, as described in sections 2.2.6 p. 79 and 2.2.8.1 p. 87 respectively, using the primers in **Table 2.2**.

5.3 Results

5.3.1 Validation and cosegregation of variants

Sanger sequencing was performed to validate variants identified through next-generation sequencing of two patients with severe early onset epilepsy, and to determine whether the variants cosegregated as expected with the parents of the patients.

5.3.1.1 Analysis of patient OTH 12

In patient OTH 12 three variants that were identified through whole genome sequencing were validated using Sanger sequencing:

1. A *de novo* c.4367G>A (p.Ser1456Asn) mutation in *C5orf42*.
2. Compound heterozygous mutations in *SCN4A*; c.2824G>A (p.Asp942Asn) and c.5424G>A (p.Met1808Ile).
3. A homozygous c.690-2A>G substitution in the *PIGQ* gene.

5.3.1.1.1 C5orf42 variant

In patient OTH 12 the heterozygous variant in the *C5orf42* gene was validated using PCR with the C5orf42-F and C5orf42-R primers described in section 2.2.6.2 (**Table 2.2**) p. 82, followed by Sanger sequencing of the PCR product using both forward and reverse primers. Sanger sequencing also showed both parents OTH 13 and OTH 14 to be homozygous for the wild type allele (**Figure 5.3**). This indicates that the variant is probably a *de novo* mutation. This was supported by analysis of the NGS data on Integrative Genomics Viewer software (Robinson *et*

al. 2011) which showed that the mutation had occurred on the paternally inherited chromosome, as paired end reads which contained the wild type allele also detected a nearby SNP that was only present in the mother (**Figure 5.4**).

5.3.1.2 SCN4A variants

The two heterozygous *SCN4A* variants were amplified using PCR with the D942N-F, D942N-R, M1809I-F, and M1809I-R primers described in section 2.2.6.2 (**Table 2.2**) p. 82, followed by Sanger sequencing of the PCR products, using both forward and reverse primers in each case. The sequencing confirmed that the patient was heterozygous for both variants, and that both parents were carriers of one variant each. However, sequencing of the unaffected sibling showed they also had both variants, suggesting that these variants are unlikely to be pathogenic (**Figure 5.5**).

5.3.1.3 PIGQ variant

Sanger sequencing was performed in patient OTH 12 and parents OTH 13 and OTH 14 using the PIGQ-F and PIGQ-R primers as described in section 2.2.6.2 (**Table 2.2**) p. 82. The resulting electropherogram confirmed that the patient was homozygous for the variant (**Figure 5.6**) and also showed that both unaffected parents and the unaffected sibling were heterozygous carriers, whereas a healthy unrelated control was homozygous for the wild type allele (**Figure 5.6**).

These results are consistent with the variant segregating as an autosomal recessive disorder. The mutation substitutes the G nucleotide of the 3' AG essential splice donor site, which would be predicted to affect splicing and possibly cause exon three to be spliced out of the mature mRNA transcript. Exon three is located immediately before a catalytic domain.

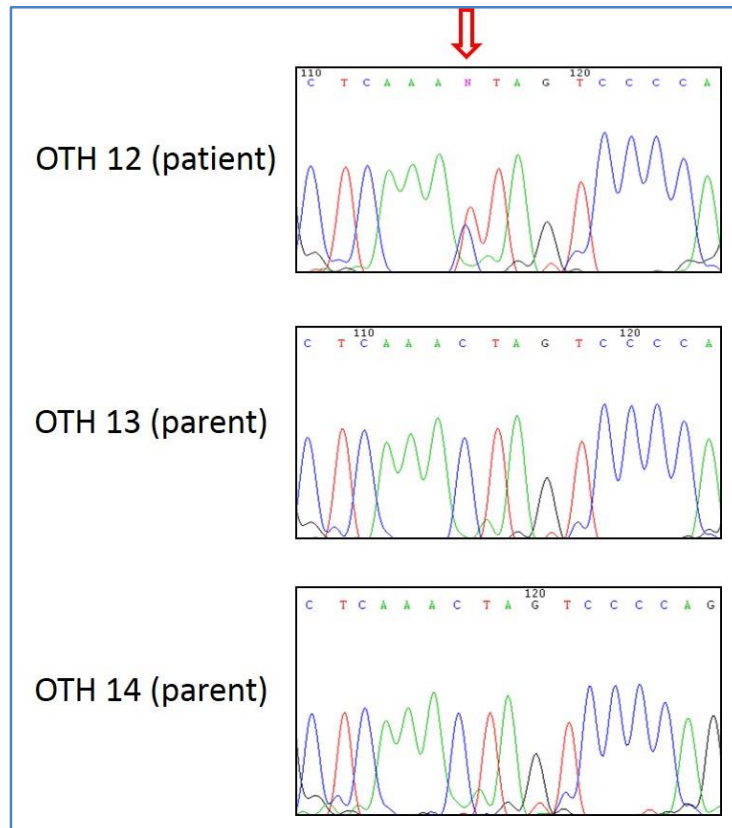


Figure 5.3 Confirmation of a heterozygous c.4367C>T substitution in the *C5orf42* gene in patient OTH 12. The red arrow shows the location of the variant in electropherograms produced by Sanger sequencing.

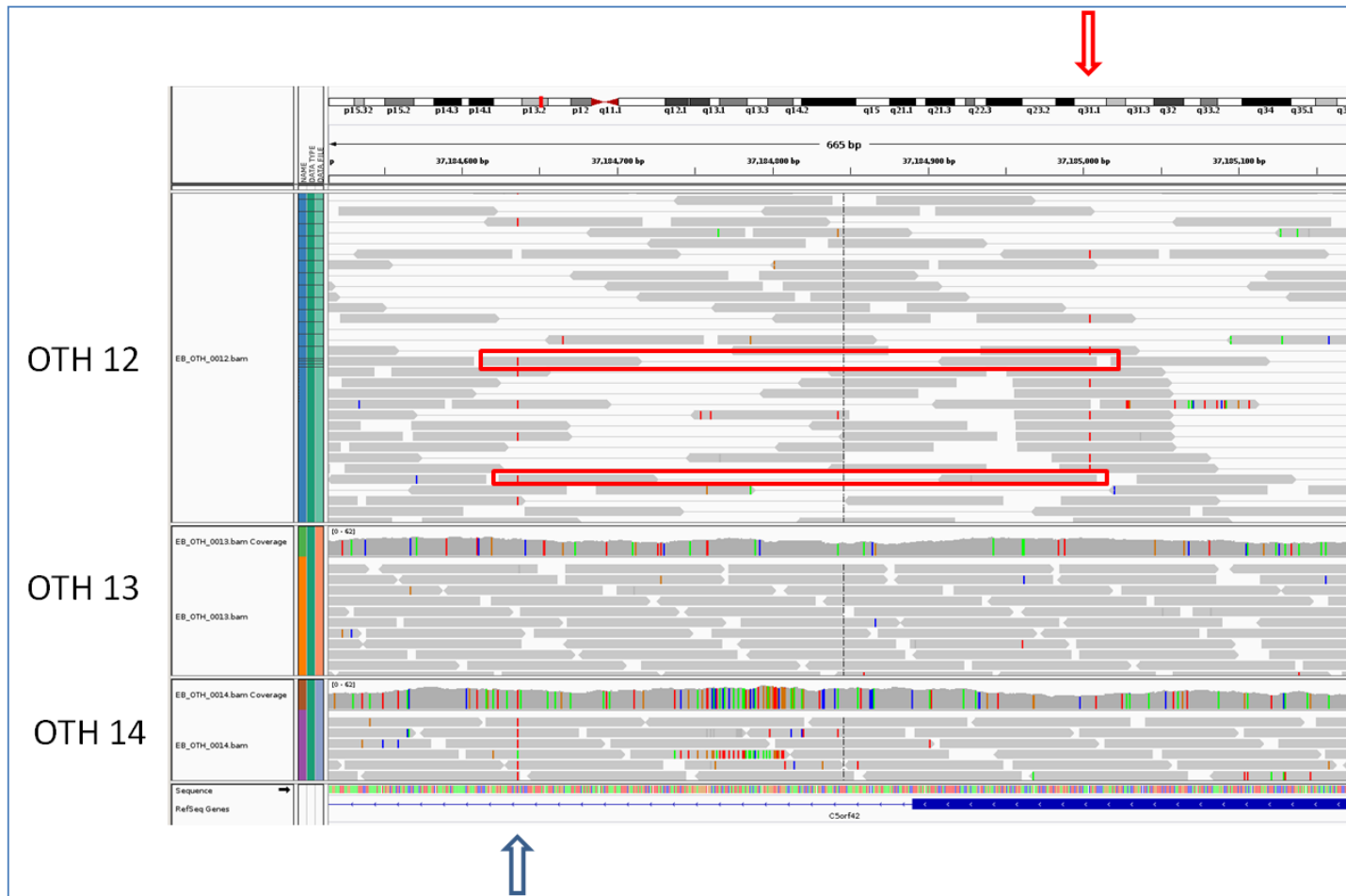


Figure 5.4 Identifying the phase of a variant using IGV. Screenshot of Integrative Genomics Viewer software showing NGS reads from the trio. The *de novo* *C5orf42* c.4367C>T substitution is indicated by the red arrow on top and the red mark in each of the grey sequence reads. The blue arrow shows a benign SNP nearby in the mother's sequence data. Two of the paired end reads show that the c.4367C>T variant does not appear in the mother's reads.

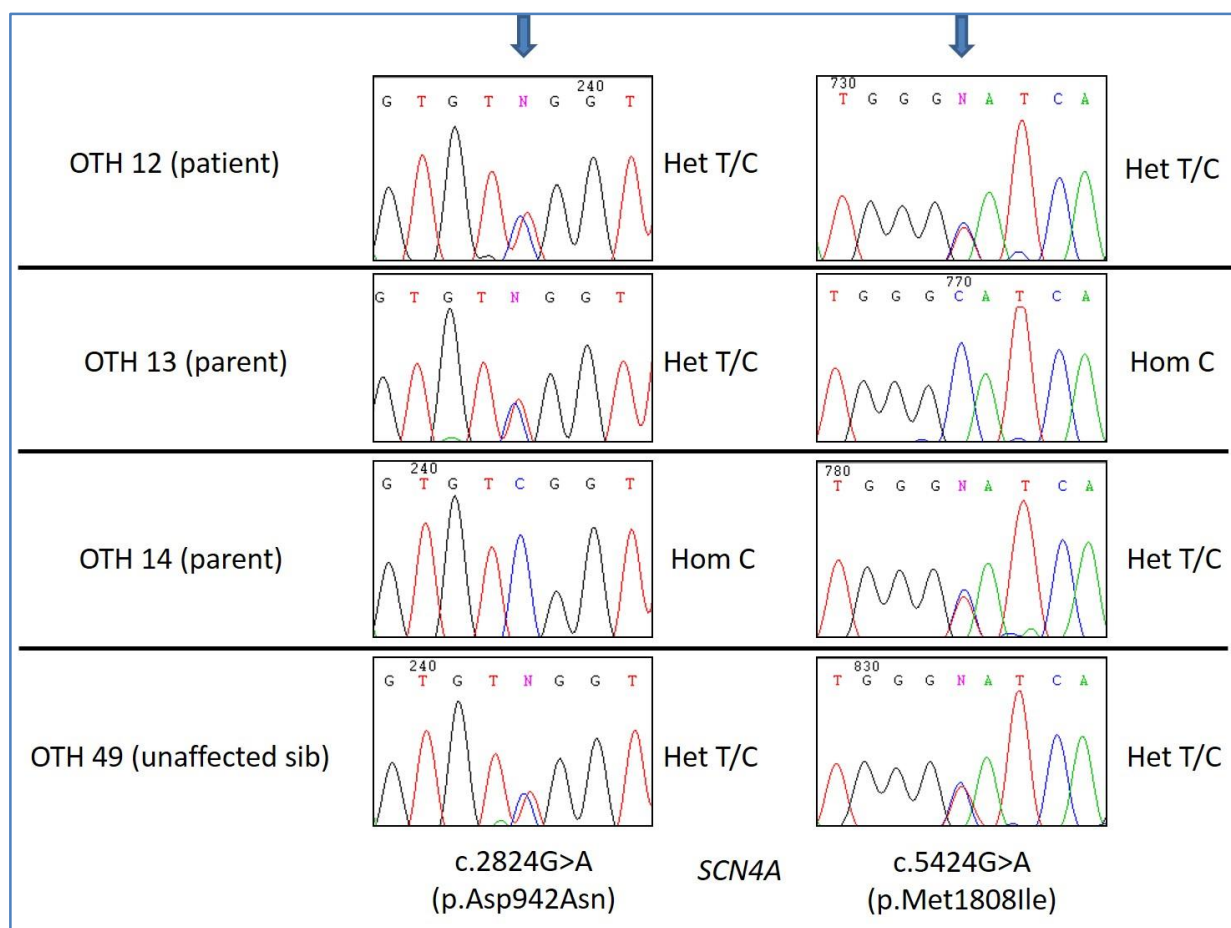


Figure 5.5 Segregation of the *SCN4A* variants in the family of OTH 12. Electropherograms produced by Sanger sequencing show that the two compound heterozygote mutations in *SCN4A* are also present in an unaffected sibling. Het = heterozygous, Hom = homozygous.

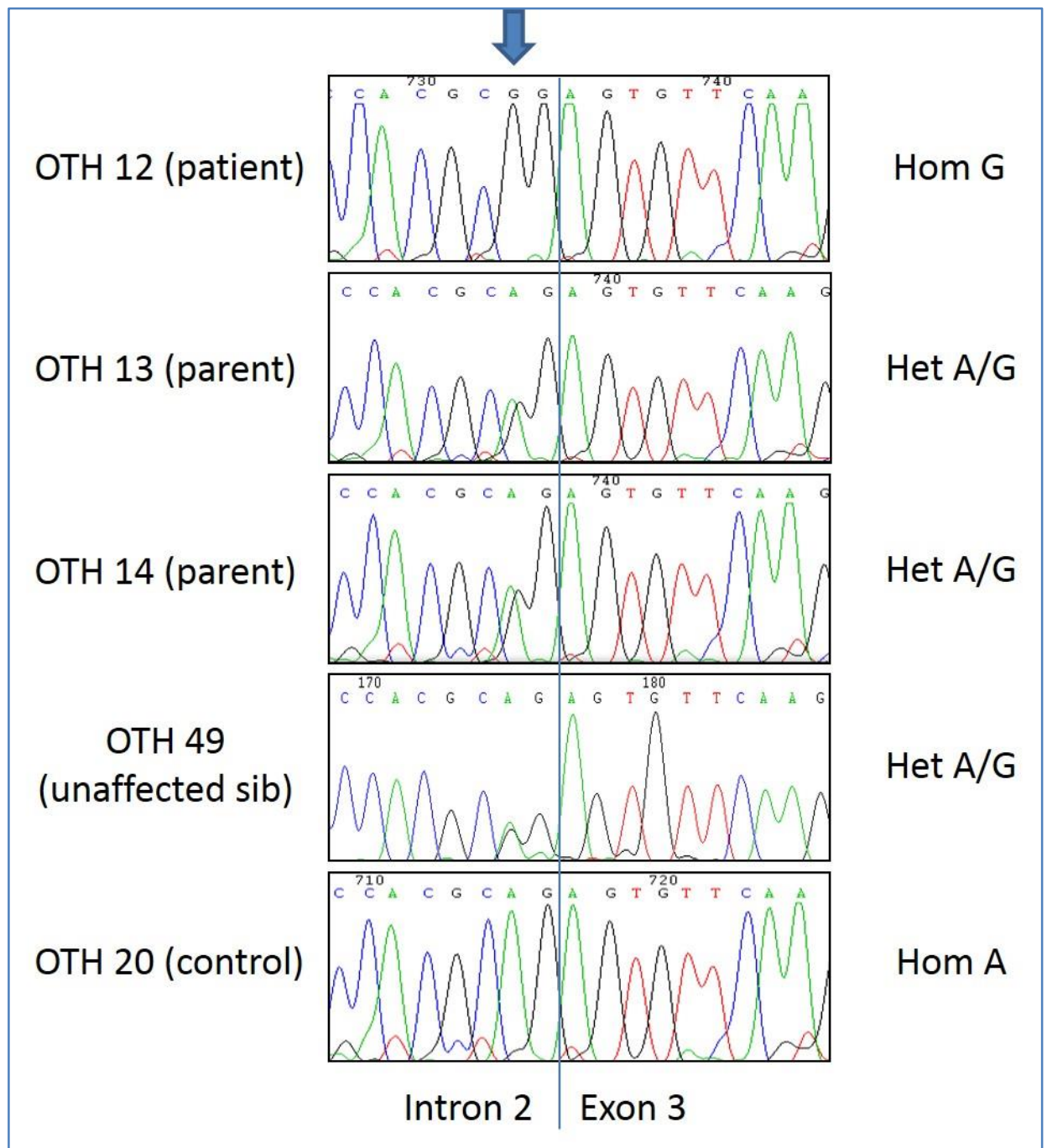


Figure 5.6 Segregation of the exon 3 c.690-2A>G substitution in the *PIGQ* gene of the family of patient OTH 12. Electropherograms from Sanger sequencing show that the variant is homozygous in the patient, heterozygous in both parents and an unaffected sibling, and absent in a control, confirming the variant segregates with the disease in an autosomal recessive inheritance pattern. Het = heterozygous, Hom = homozygous.

5.3.1.2 Patient OTH 18

In patient OTH 18 a c.1228-1G>A in intron 8 of the *CBL* gene had been identified which was predicted to affect splicing. Analysis of the NGS reads using Integrative Genomics Viewer showed another nearby SNV, a c.1321C>T substitution which is synonymous, and so likely to be benign (**Figure 5.7**). In the NGS data it appeared that both of these SNVs were *de novo* as neither of the parents showed either mutation in each of those positions (**Figure 5.7**). Both variants in the *CBL* gene were validated by amplification using the CBL-F and CBL-R primers described in section 2.2.6.2 (**Table 2.2**) p. 82 followed by Sanger sequencing using both the forward and reverse primers.

Sanger sequencing was also performed to confirm that the variants were absent in the parents (**Figure 5.8**). Parental testing performed by the NHS confirmed that OTH 19 and OTH 20 were the parents of OTH 18. Both of these variants were therefore *de novo* changes in the patient.

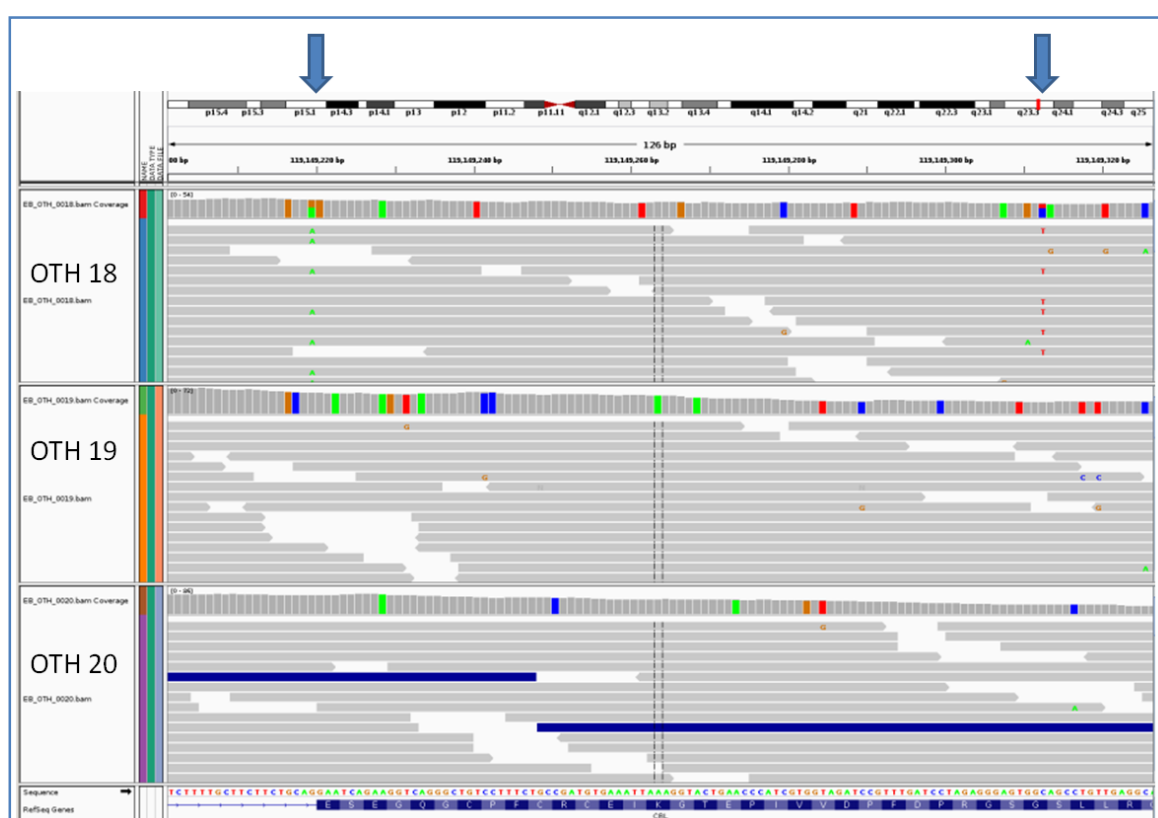


Figure 5.7 NGS data showing two *de novo* SNVs in patient OTH 18. The positions of the variants are indicated by blue arrows, and by green and red marks on the grey sequencing reads. The SNV on the right is c.1228-1G>A.

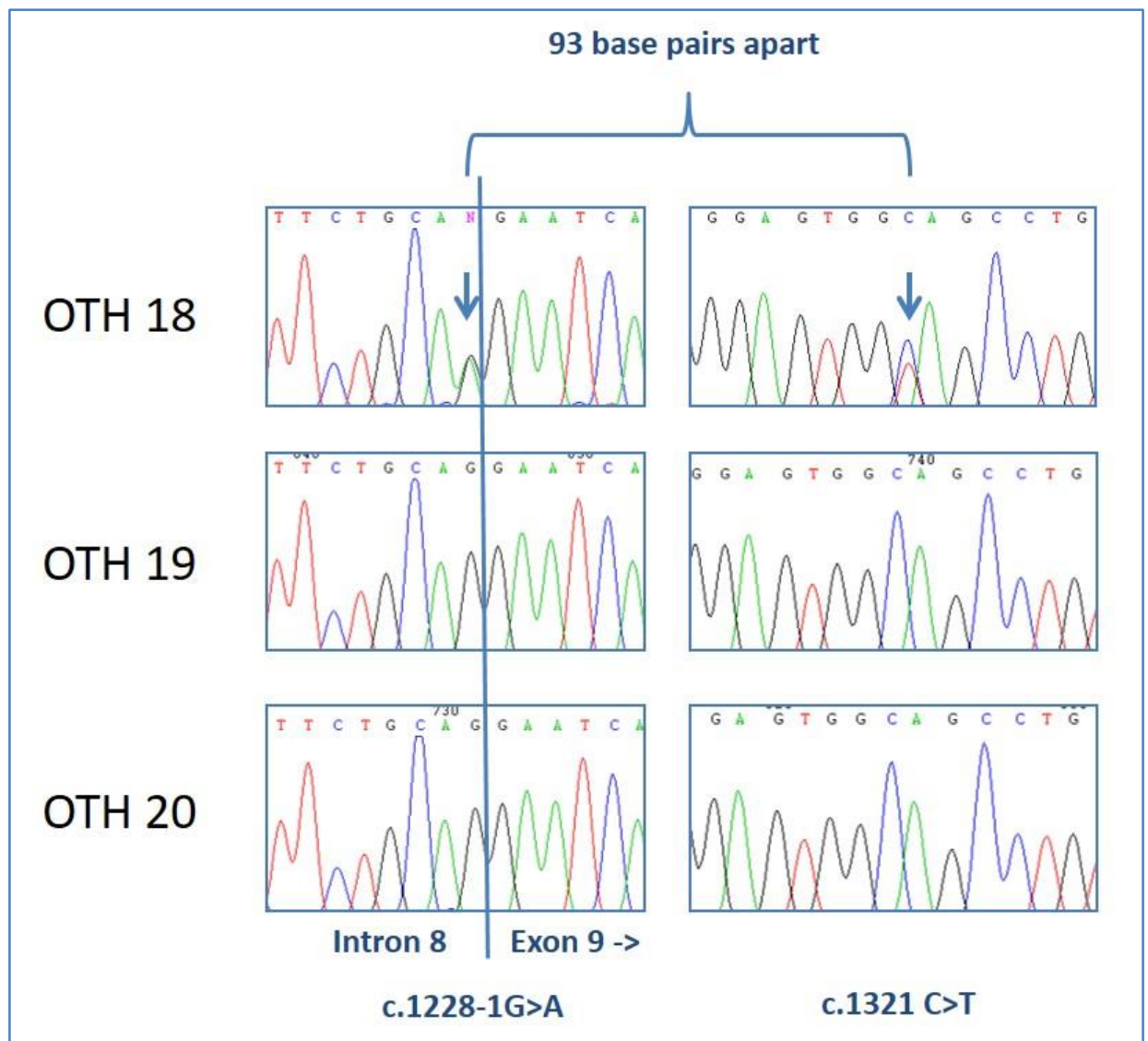


Figure 5.8 Two SNVs in the *CBL* gene of patient OTH 18. Electropherograms show that the two SNVs in the *CBL* gene of patient OTH 18 were *de novo* mutations in the patient, as they do not appear in the two parents; OTH 19, and OTH 20. The SNV on the left is c.1228-1G>A which is located on the final base of the intron and mutates the guanine base of the AG splice acceptor site.

5.3.2 Determination of splicing effect in mRNA

The exon 3 c.690-2A>G substitution in the *PIGQ* gene of patient OTH 12 and the c.1228-1G>A substitution in the *CBL* gene of patient OTH 18 both mutate one of the bases in the essential splice acceptor site of the intron, which may cause skipping of the next exon. Having validated these variants in genomic DNA using Sanger sequencing and confirmed segregation of the variants with the disease in family members, the next investigation was to determine if they were actually having an effect on splicing.

A literature search showed that both genes have been reported to be ubiquitously expressed. Therefore, mRNA was extracted from patient and parent lymphocytes. The extracted mRNA was reverse transcribed into cDNA, which was then amplified by PCR and run on an Agilent 2100 Bioanalyzer with DNA 7500 chip and reagents (sections 5.2.1 and 5.2.2 pp 173 and 174).

The families were tested for both of the mutations, allowing the other family to be used as a control in each case.

5.3.2.1 Patient OTH 12

The patient OTH 12 had died before it was possible to obtain a blood sample to perform mRNA experiments, so only the parents OTH 13 and OTH 14 could be tested. Primers were designed to amplify cDNA across exons 2-4 of the *PIGQ* gene. The amplification products produced were analysed on an Agilent 2100 Bioanalyzer (section 2.2.5 p. 78), which showed the expected band at 583bp indicating a product from a transcript which contains exon 3, and a lower band of 451 base pairs which suggests the presence of a transcript that skips the exon (exon 3 is 132 base pairs in size), but does not cause a frameshift (**Figure 5.9**). The other samples show just one band of 583bp in size, produced from a single mRNA transcript of the wild type sequence containing exon 3.

These products were Sanger sequenced, which confirmed that the product is missing exon 3, indicating that the presence of the c.690-2A>G variant causes skipping of exon 3 (**Figure 5.10**).

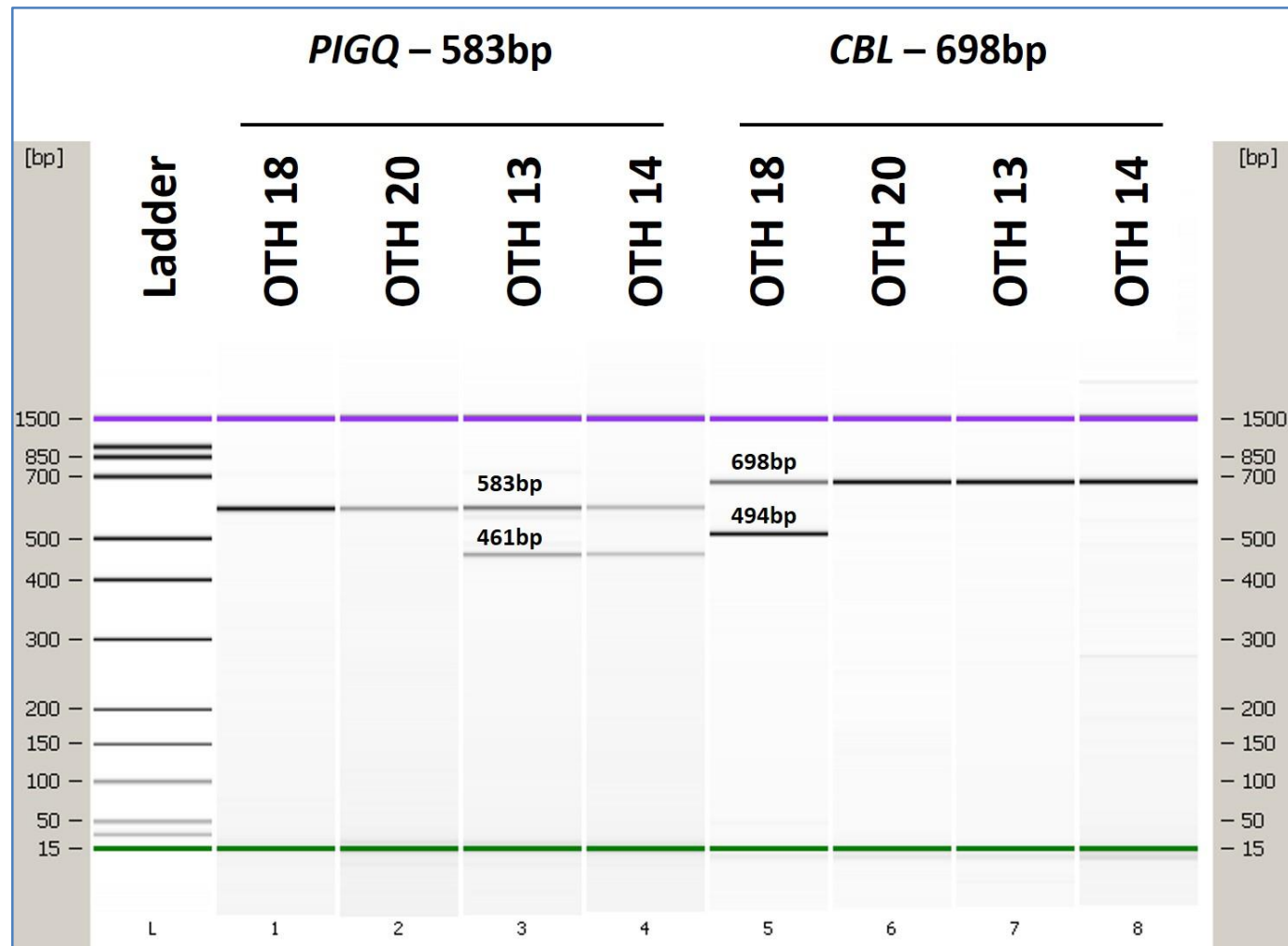


Figure 5.9 Bioanalyzer data showing amplicons of cDNA from samples OTH 18, 20, 13, and 14. The cDNA was amplified for regions covering both the *CBL* and *PIGQ* gene splice mutations. OTH 13 and OTH 14 both produce a lower band of 461bp in size that does not contain exon three of the *PIGQ* gene. OTH 18 is heterozygous for the *CBL* variant which causes exon nine to be skipped, generating a lower band of 494bp in size.

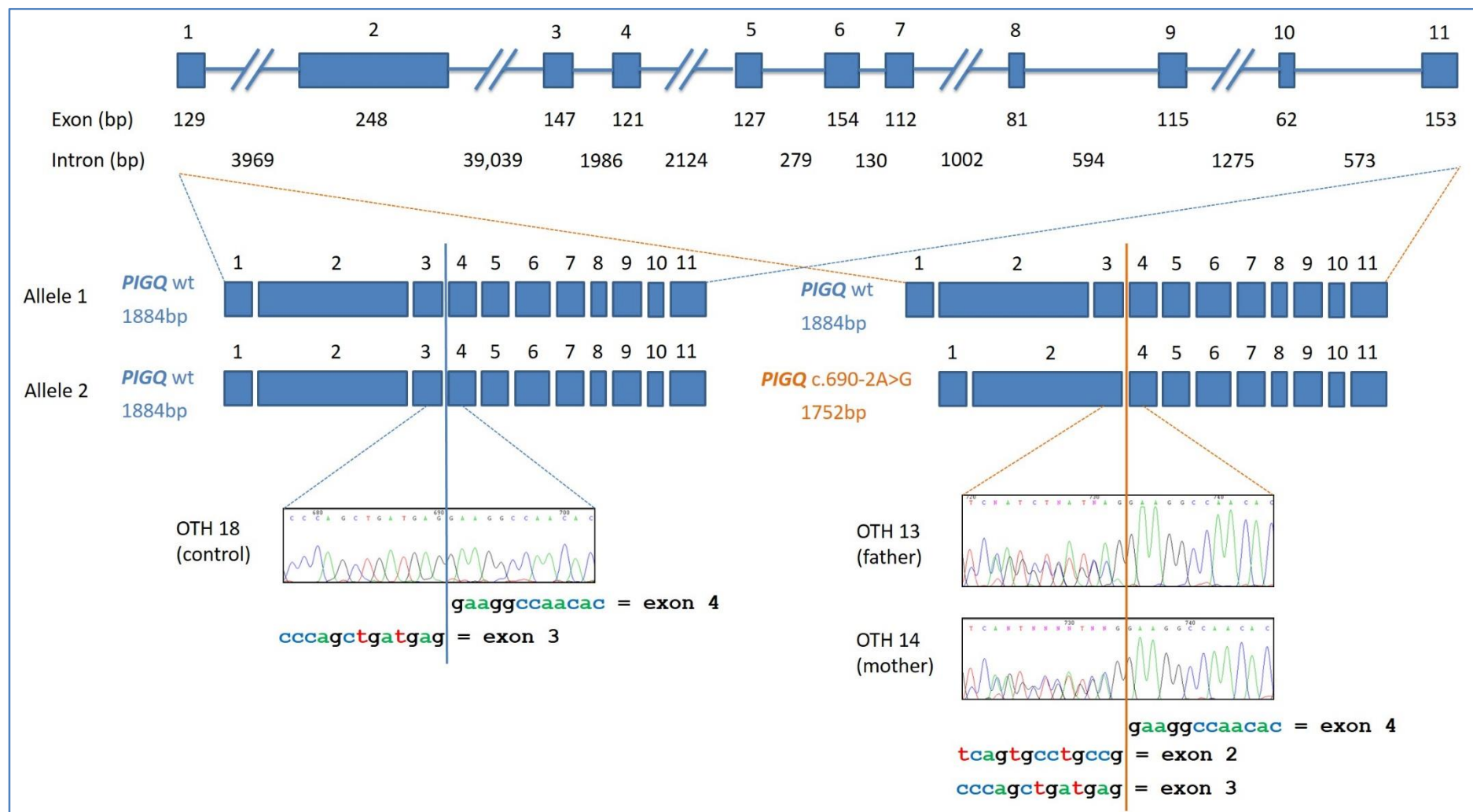


Figure 5.10 Exon skipping in the *PIGQ* gene in the parents of patient OTH 12. Schematic diagram and electropherograms produced by Sanger sequencing of cDNA showing the skipped exon three in the mRNA transcript produced by the c.690-2 A>G mutation in *PIGQ* in one allele of each of the parents of patient OTH 12.

5.3.2.2 Patient OTH 18

In patient OTH 18 the c.1228-1G>A substitution in intron 8 of the *CBL* gene was predicted to cause skipping of exon 9, which is 204bp in size, however the resulting transcript is still in frame. Primers were designed to cover exons 8 to 10. The Bioanalyzer trace of these amplifications in patient OTH 18 showed two bands for the *CBL* amplicon, one at 698bp as expected, and another at 494bp, suggesting that the variant causes exon 9 to be skipped (**Figure 5.9**). The other three samples have just one band from the wild type sequence.

This was confirmed by Sanger sequencing of the PCR products (**Figure 5.11**). In the unaffected father OTH 20, a single trace is visible with the end of exon 8 leading into exon 9 at the blue line. In the affected patient OTH 18, the trace instead shows two sequences after the gold line; exon 9, and a faint trace showing exon 10, indicating that on one allele exon 9 of the *CBL* gene has been skipped.

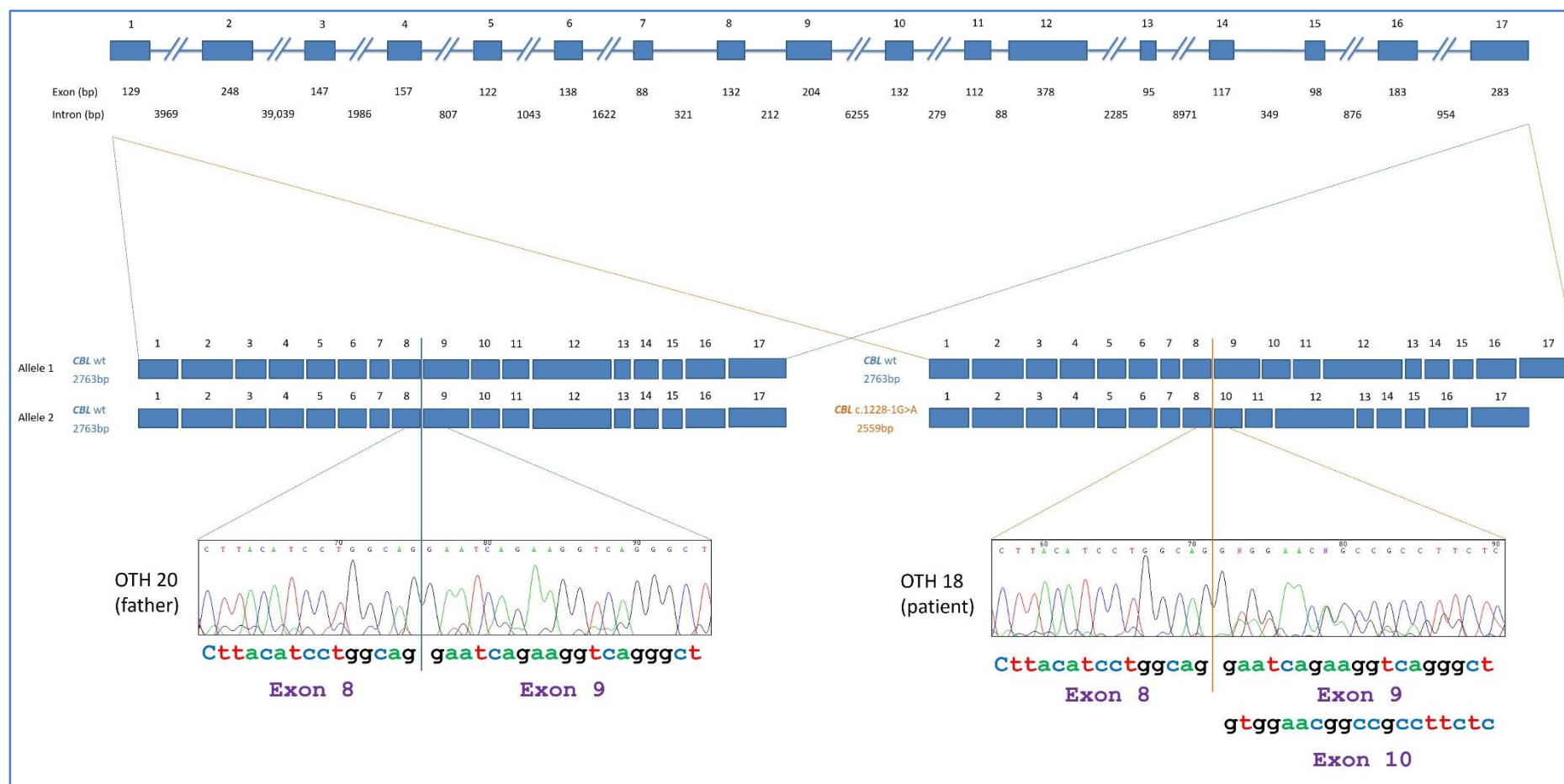


Figure 5.11 Exon skipping in the *CBL* gene in patient OTH 18. Schematic diagram and electropherograms produced by Sanger sequencing of cDNA showing the skipped exon nine in the mRNA transcript produced by the c.1228-1 G>A mutation in *CBL* on one allele of patient OTH 18.

5.4 Discussion

This chapter describes investigations into two probands (OTH 12 and OTH 18) with early-onset epilepsy in whom four potentially pathogenic variants had been identified through whole genome sequencing (section 5.1.2.1 p. 169).

5.4.1 Mutations in *PIGQ* can cause Ohtahara syndrome

A c.690-2A>G substitution was identified in the *PIGQ* gene of patient OTH 12 who had been diagnosed with Ohtahara syndrome and severe developmental delay. The variant was selected from the whole genome sequencing data as a potentially pathogenic variant because the protein it encodes forms part of a complex that catalyses the first step of the GPI biosynthesis pathway (Almeida *et al.* 2009). Mutations in several other genes involved in this pathway have been reported to cause Ohtahara syndrome, as shown in **Figure 5.12** (Almeida *et al.* 2006, Krawitz *et al.* 2010, Maydan *et al.* 2011, Johnston *et al.* 2012).

In addition, the phenotype of patient OTH 12 most closely matches the phenotype of patients with mutations in *PIGA* (hypotonia, seizures with burst suppression EEG, thin corpus callosum, lethal in infancy). *PIGQ* forms part of the same complex as *PIGA*, and is therefore involved in the same step of the GPI biosynthesis pathway (Johnston *et al.* 2012).

Sanger sequencing confirmed that the patient and both parents of OTH 12 were carriers of the variant, suggesting consanguinity. The parents were members of the Yoruban tribe from West Africa and the relatively small size of the tribe could have contributed to a founder effect that allowed the c.690-2A>G variant to spread.

Testing of cDNA from the parents confirmed that the c.690-2A>G variant in *PIGQ* caused skipping of exon three in the transcript produced by the gene. Exon three is located immediately before a catalytic domain, and so the change would be expected to have a functional effect on the protein produced. This was the first report that mutations in *PIGQ* can cause Ohtahara syndrome (Martin *et al.* 2014).

GPI biosynthesis pathway

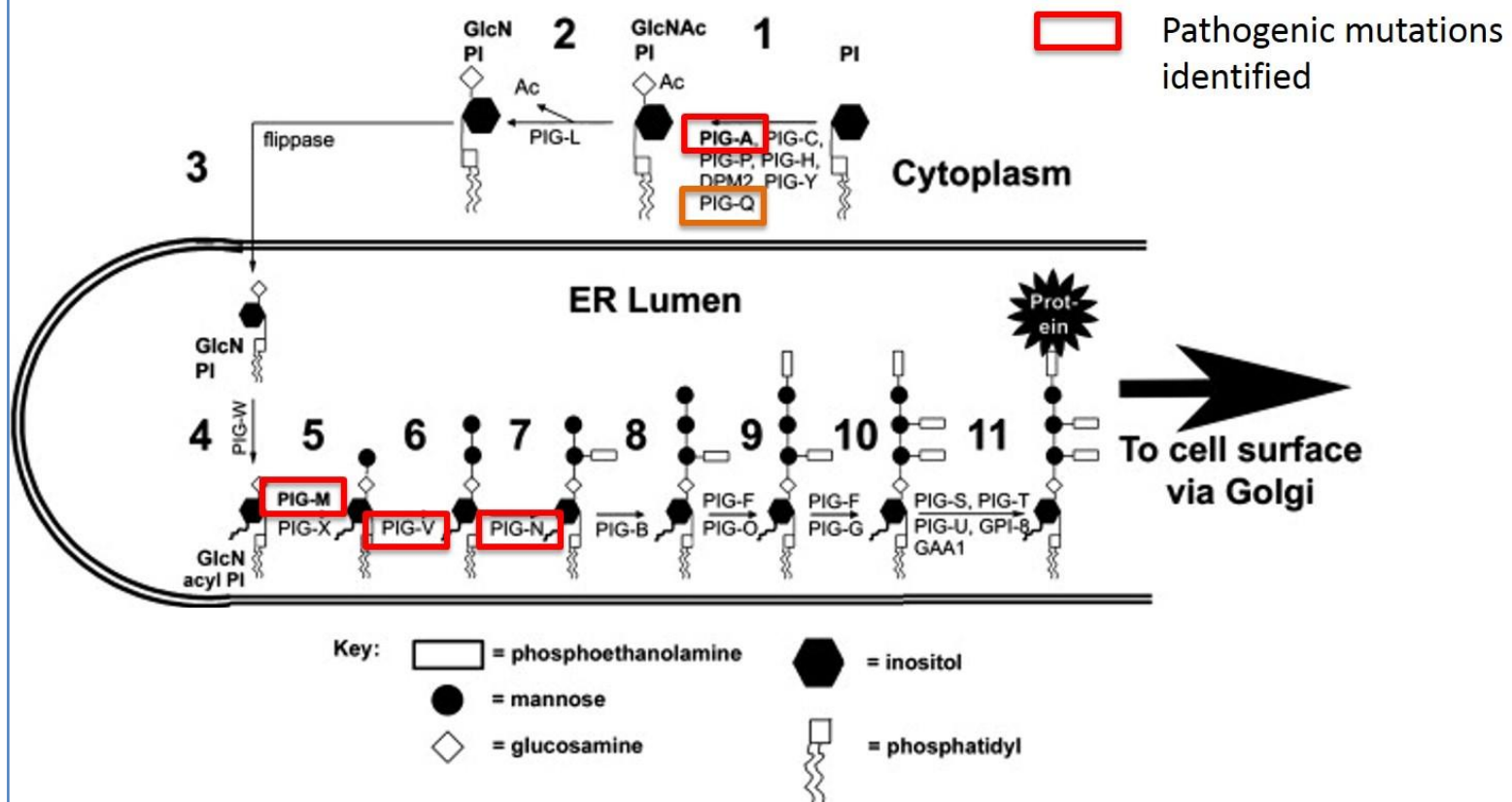


Figure 5.12 GPI biosynthesis pathway. The genes encoding the subunits highlighted in red have been shown to cause severe early epilepsy when mutated. *PIGQ* is highlighted in orange. Figure adapted from Almeida *et al.* (2009)

5.4.2 A variant in *CBL* causes a range of phenotypes

Patient OTH 18 had Severe Non-Syndromic Early Onset Epilepsy, severe developmental delay, patent ductus arteriosus, and atrial septal defect as a neonate, as well as cutaneous hypopigmentation. A heterozygous c.1228-1G>A substitution in the *CBL* gene was identified by whole genome sequencing.

The variant was confirmed as being *de novo*, as was c.1321C>T, another nearby synonymous SNV of unknown significance. The presence of multiple *de novo* mutations in close proximity has previously been reported by O’Roak and colleagues in cases of sporadic autism, and supports our current understanding of how mutations occur; if a region of the genome is exposed to a mutagenic effect multiple mutations may occur in close proximity (O’Roak *et al.* 2012).

The CBL protein is named after Casitas B–lineage Lymphoma, and it ubiquitinates receptor tyrosine kinases involved in the Ras/MAPK pathway (Swaminathan and Tsygankov 2006), which is most frequently associated with cancer, rather than epilepsy. However, mutations in two other genes involved in the Ras/MAPK pathway; Kirsten Rat Sarcoma Viral Oncogene Homolog (*KRAS*), and V-RAF Murine Sarcoma Viral Oncogene Homolog B (*BRAF*), have been reported to cause cardiofaciocutaneous syndrome (CFCS) with refractory epilepsy (Adachi *et al.* 2012) and also Noonan syndrome (van der Burgt 2007). Both CFCS and Noonan syndrome are heterogeneous disorders that show overlap in phenotypes, which include developmental delay, short stature, congenital heart defect, broad or webbed neck, and characteristic facial dysmorphisms.

Patient OTH 18 did not show the typical appearance of CFCS, however the patient did have a large hypo-pigmented patch, as well as an atrial septal defect as an infant, suggesting the patient might have a CFCS-like syndrome. The difference in phenotype would be expected as the *CBL* gene is involved in another step of the Ras/MAPK pathway, and mutations in the gene have been reported to show a great deal of phenotypic heterogeneity (Perez *et al.* 2010).

Confirmation that this variant caused the skipping of exon 9 in OTH 18 was performed using mRNA extracted from blood. The deletion of this exon would be expected to affect the ubiquitin ligase activity of *CBL* as this exon codes for the C-terminal part of the RING finger domain which causes that activity, and would have an ultimate effect on the Ras/MAPK pathway, which lends further evidence to suggest that the *CBL* variant is the prime candidate to be the cause of the patient's condition.

5.4.3 Conclusion

The work described in this chapter has provided evidence to support the hypothesis that two variants identified by whole genome sequencing of two patients with severe early onset epilepsy were the cause of the condition in these patients.

In the case of patient OTH 12 these results are the first to report that mutations in *PIGQ* can cause severe early onset epilepsy. It seems plausible that mutations in the genes coding for the remaining parts of the GPI biosynthesis pathway could also cause Ohtahara syndrome, or a related early onset epilepsy phenotype, and in future work these genes should be screened in patients who present with a similar phenotype.

In the case of patient OTH 18 the variant identified in *CBL* remains the prime candidate, even though the phenotype of the patient does not fit with CFCS. If the *CBL* variant is the cause of the condition it would suggest that mutations in this gene can give rise to a greater range of clinical phenotypes than previously thought. This suggests that in future work a broader range of clinical phenotypes should be screened for mutations in *CBL*, to more fully characterise the spectrum of clinical phenotypes which mutations in this gene can cause.

This work demonstrates the need to perform further experiments other than sequencing to reach a secure molecular diagnosis for patients with genetic diseases. Testing cosegregation of disease variants in family members can usually be performed in clinic, however techniques required to provide further evidence such as testing of mRNA are not always routinely available

in clinical diagnostics labs. Further testing such as functional assays of the activity of mutant proteins compared to the wild type are even more likely to be restricted to a research setting.

This work has shown the utility of WGS in clinical practice. Both of these patients were selected for WGS as they had previously been refractory to molecular diagnosis. At the time of the work associations between genes and epilepsy were still being discovered, which would have made designing a comprehensive capture design for targeted sequencing impossible. Exome sequencing could have been used, however it lacks the ability to identify novel genes, which any type of targeted capture by definition cannot do.

Data from the investigations described in this chapter were published in *Human Molecular Genetics* (Martin *et al.* 2014).

CHAPTER 6

Searching for pathogenic variants in macular dystrophies

6.1 Introduction

Macular dystrophies (MD) are a clinically and genetically heterogeneous group of disorders characterised by decreased visual acuity and loss of colour vision, which caused deterioration of the macula (the central region of the retina) as introduced in section 1.8.1 p. 54. Macular dystrophies were classically characterised as conditions with predominantly cone loss, but many macular dystrophies have both cone and rod involvement, and can more accurately be classified as cone-rod dystrophies.

The most common macular dystrophy is Stargardt disease (STGD1, OMIM #248200). The clinical features first described by German ophthalmologist Karl Stargardt (1909) typically comprised early onset macular flecks with varying degrees of atrophy of the macula. The atrophy usually occurs as the disease progresses, starting in the macula as shown in **Figure 6.1 A**. Not all cases have typical flecks at presentation, but these can develop over time. **Figure 6.1 B** shows a case with typical flecks surrounding developing central atrophy. The phenotype can range from patches of atrophy which coalesce over time, as shown in **Figure 6.1 C i**, to maculopathies with a bull's eye appearance, as shown in **Figure 6.1 D i** and **Figure 6.2** which are clearly illustrated using other imaging modalities in **Figures 6.1 C ii and D ii** as described in the figure legend. The age of onset is typically in childhood or early adolescence and is inherited as an autosomal recessive inheritance pattern, and an estimated incidence rate of 1:10,000 (Maugeri *et al.* 1999).

Although originally classified as a macular dystrophy, Stargardt disease shows a great deal of heterogeneity, and once electrophysiology and genetic testing became available for characterising the condition it became apparent that not all patients had a condition confined to the macula. Indeed, in a proportion of patients diagnosed with Stargardt disease a more severe degeneration occurs, such as chorioretinal atrophy, shown in **Figure 6.3**.

Fundus flavimaculatus (FF) is a descriptive term used to denote the more widespread presence of white or yellow flecks throughout the retina, usually sparing the far periphery, in addition to

areas of macular atrophy that often resemble Stargardt disease (Franceschetti and Francois 1965). An example of fundus flavimaculatus is shown in **Figure 6.4**.

The revolution in molecular genetic techniques in the 1980's and 1990's allowed linkage analysis to be performed in an attempt to determine the underlying genetic basis of Stargardt disease and related disorders. The phenotype was defined as clearly as possible to ensure a likely genetically homogeneous condition was being investigated.

The gene associated with Stargardt disease was subsequently mapped to the short arm of chromosome 1 (Kaplan *et al.* 1993) and further linkage analysis suggested that fundus flavimaculatus (FF) also mapped to the same region, supporting earlier hypotheses that these conditions were allelic disorders (Gerber *et al.* 1995). Fine mapping more accurately mapped the gene to a 4-cM interval on 1p21-22 (Hoyng *et al.* 1996).

In 1997 mutations in the gene now known as *ABCA4* (ATP-binding cassette transporter, alpha 4 subunit, originally called *ABCR*) were identified in patients with STGD1 (Allikmets *et al.* 1997b). In the original paper 36 exons of the gene were reported, but this number was subsequently revised to 50 exons in a further paper in which the complete exon-intron structure of the gene was described, as shown in **Figure 6.5** (Gerber *et al.* 1998).

ABCA4 encodes the alpha 4 subunit of the ATP binding cassette transporter superfamily, a group of transmembrane proteins that transport molecules across membranes (Azarian and Travis 1997, Illing *et al.* 1997). The protein encoded by the gene is localised to the rim of the disc membrane of retinal rod outer segments (Sun and Nathans 1997). The substrates of the *ABCA4* rim protein transporter are phosphatidylethanolamine (PE) and N-retinylidene phosphatidylethanolamine (NRPE). Mutations in *ABCA4* cause the build-up of these toxic compounds which in turn causes photoreceptor death (Weng *et al.* 1999, Quazi *et al.* 2012). The structure of the gene transcript and the protein structure are shown in **Figure 6.6**.

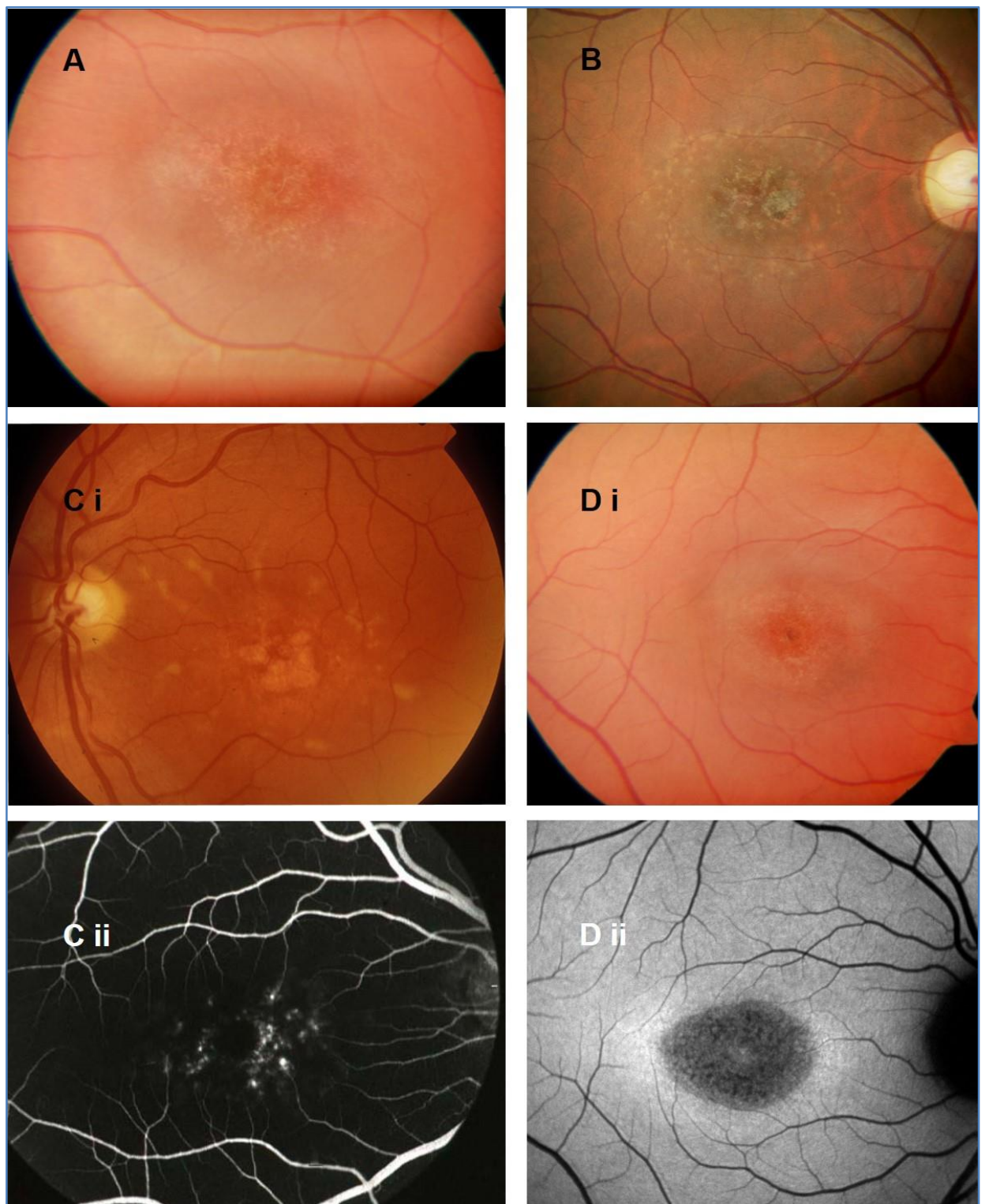


Figure 6.1 Images of Stargardt disease in four patients. A) Patchy early atrophic change. B) Typical flecks surrounding atrophic area. C) i) patchy atrophic change ii) One method used in the past for diagnosis of STGD was fluorescein angiography. Here is a picture with typical dark choroid. D) i) Bulls eye dystrophy. ii) The atrophy shows more clearly on autofluorescence imaging.

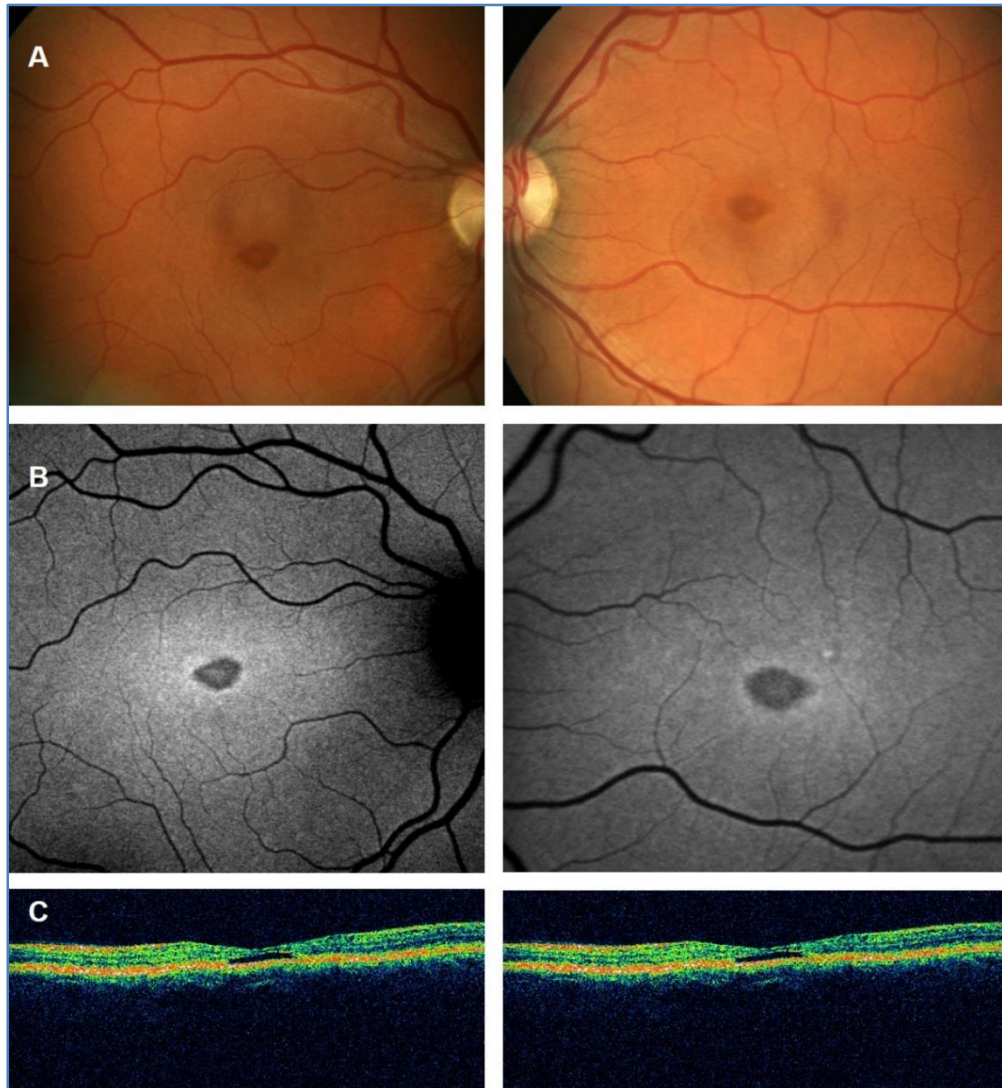


Figure 6.2 Various images of a bullseye maculopathy. A) Colour photograph showing central bullseye appearance. B) Imaging shows decreased central signal with increased ring of autofluorescence. C) Optical coherence tomography showing central abnormality.

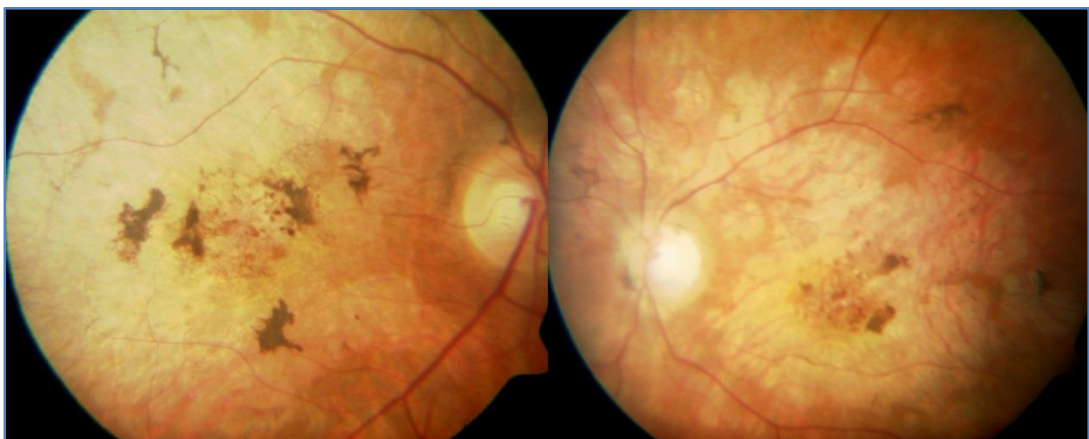


Figure 6.3 Colour photograph showing chorioretinal atrophy. In some cases STGD progresses into a more severe condition, such as the chorioretinal atrophy shown in this image.

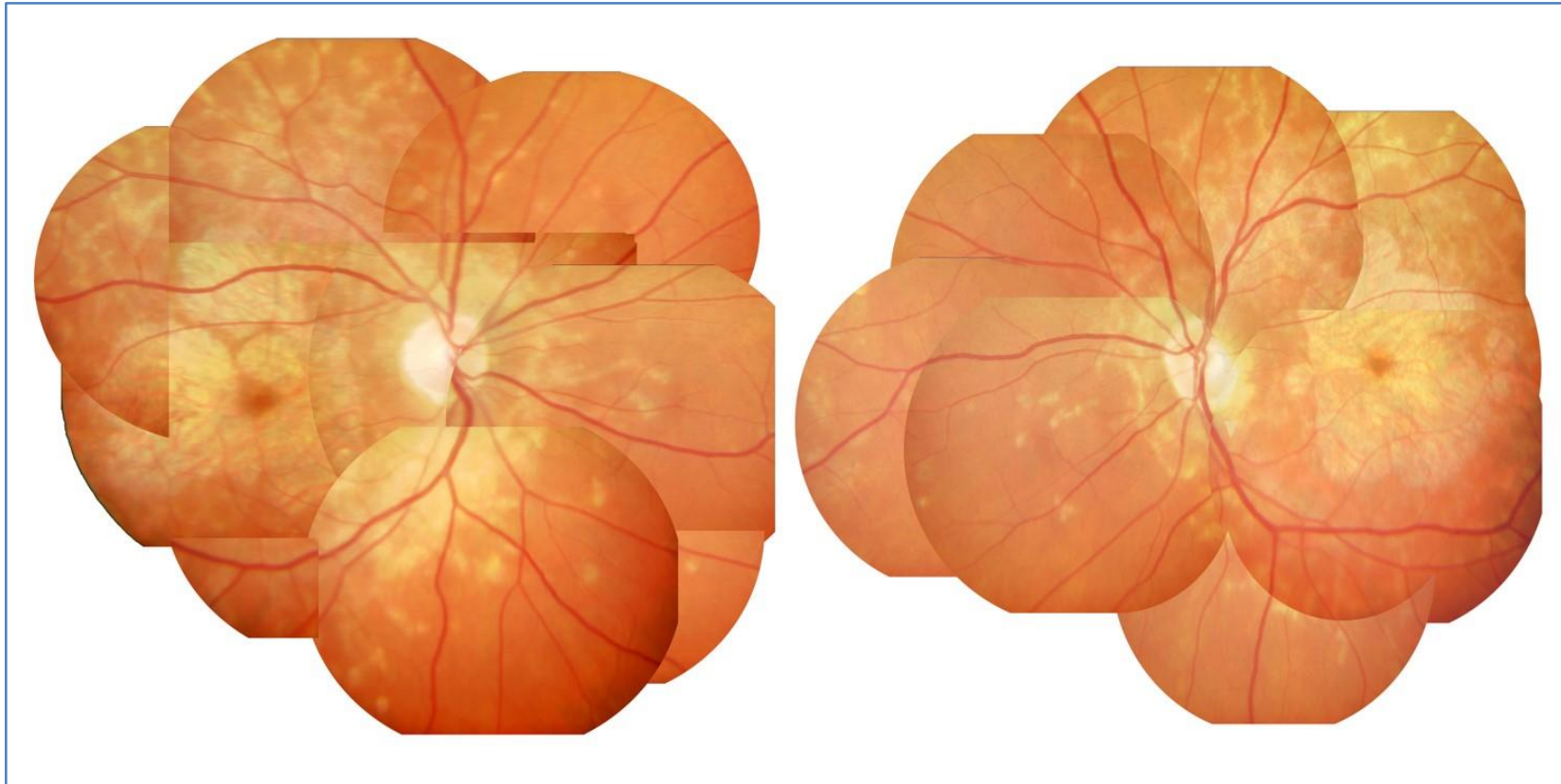


Figure 6.4 Composite image of a patient showing Fundus Flavimaculatus.

Once mutations in *ABCA4* were identified, subsequent analysis of other macular dystrophies revealed that some patients with a phenotype that was not typical of Stargardt disease also had mutations in *ABCA4* (Lewis *et al.* 1999, Rozet *et al.* 1999, Molday *et al.* 2000, Shroyer *et al.* 2000, Gerth *et al.* 2002). Cremers and colleagues reported splice site mutations in *ABCA4* causing both retinitis pigmentosa (RP) and cone-rod dystrophy (CRD) phenotypes (Cremers *et al.* 1998). Maugeri and colleagues suggested in 2000 that mutations in *ABCA4* are the major cause of autosomal recessive cone-rod dystrophy (Maugeri *et al.* 2000). Numerous other authors have also reported cases of RP and CRD caused by mutations in the *ABCA4* gene (Martinez-Mir *et al.* 1998, Klevering *et al.* 1999, Fishman *et al.* 2003, Klevering *et al.* 2005). In addition there were reports of heterozygous mutations in *ABCA4* increasing the risk of age related macular degeneration (AMD, Allikmets *et al.* 1997a, Allikmets 2000, Shroyer *et al.* 2001, Bernstein *et al.* 2002), so it became clear that Stargardt disease was not the only type of IRD caused by *ABCA4* mutations and the term *ABCA4* retinopathies was coined.

The effect of a pathogenic mutation in the *ABCA4* gene and consequent impact on *ABCA4* transporter function has been proposed to influence the severity of the disease phenotype. Nonsense mutations, which lead to partial or total loss of the *ABCA4* protein cause the more severe, widespread deterioration of CRD or RP (Armstrong *et al.* 1998, Shroyer *et al.* 1999, Shroyer *et al.* 2001, Yatsenko *et al.* 2001), whereas missense mutations which tend to leave the protein intact but with variable effects on function, lead to conditions with a later age of onset, such as STGD1, FF, and probably AMD, as shown in **Figure 6.7** (Maugeri *et al.* 1999).

In addition to autosomal recessive STGD1, some patients were found to have clinical phenotypes resembling STGD1, but a dominant pattern of inheritance. Genetic analysis of these phenocopies showed that they were caused by mutations in other genetic loci. STGD3 (OMIM #600110) is caused by mutations in *ELOVL4* (Zhang *et al.* 2001), and STGD4 (OMIM #603786) is caused by mutations in the *PROM1* gene (Yang *et al.* 2008), both of which have an autosomal dominant inheritance pattern.

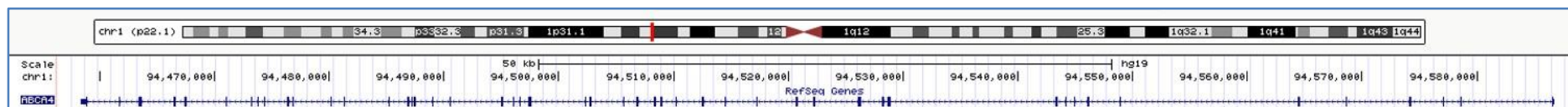


Figure 6.5 Structure of the *ABCA4* gene. The *ABCA4* gene displayed here on UCSC genome browser, contains 50 exons and is 128,312 base pairs in length. The gene is located on chromosome 1 p22.1, as indicated by the red mark at the top of the image.

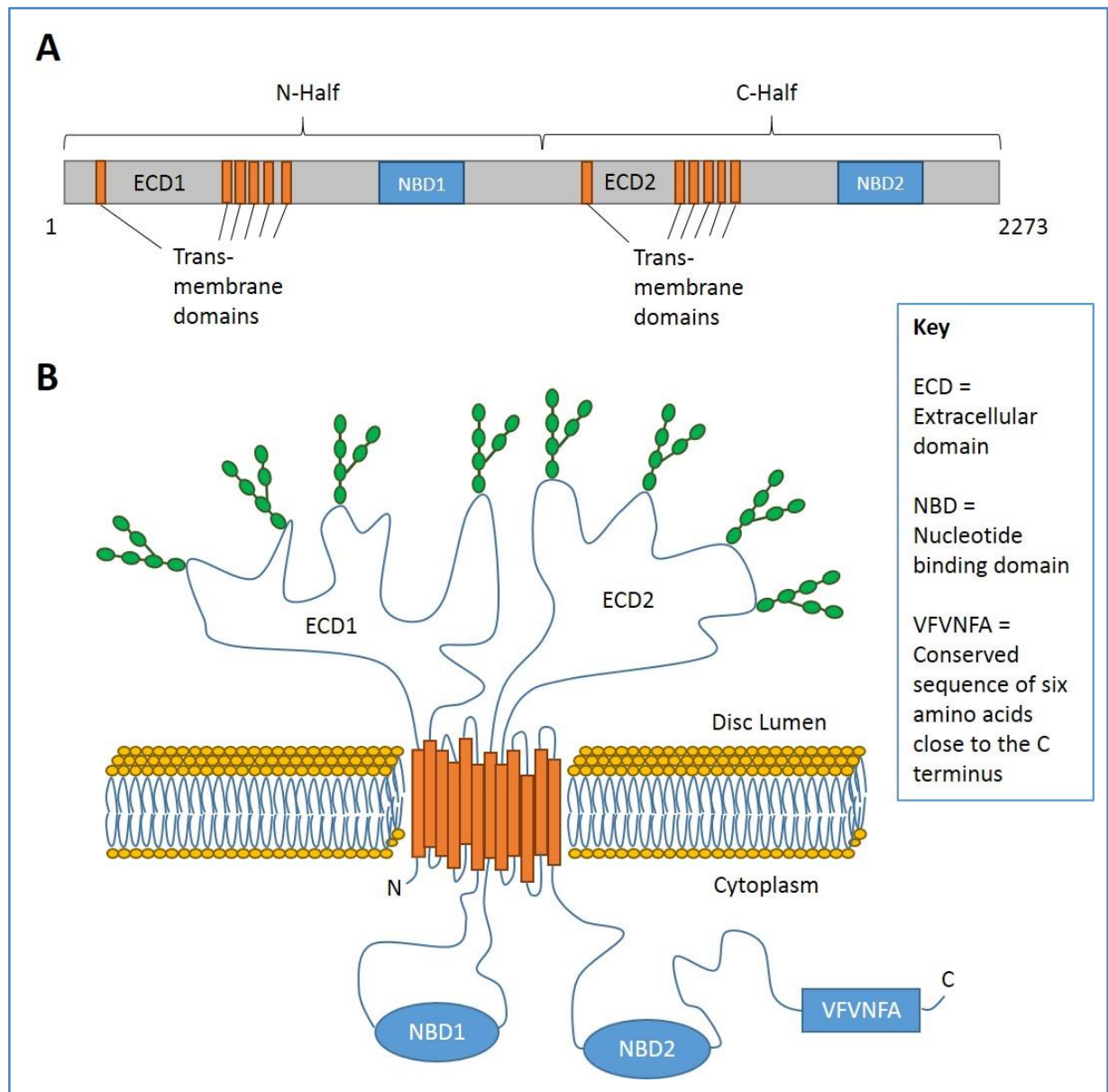


Figure 6.6 Structure of the ABCA4 transporter protein. **A)** Linear diagram of the gene transcript showing regions coding for structural elements in series. **B)** Schematic of the ABCA4 transporter protein in the membrane.

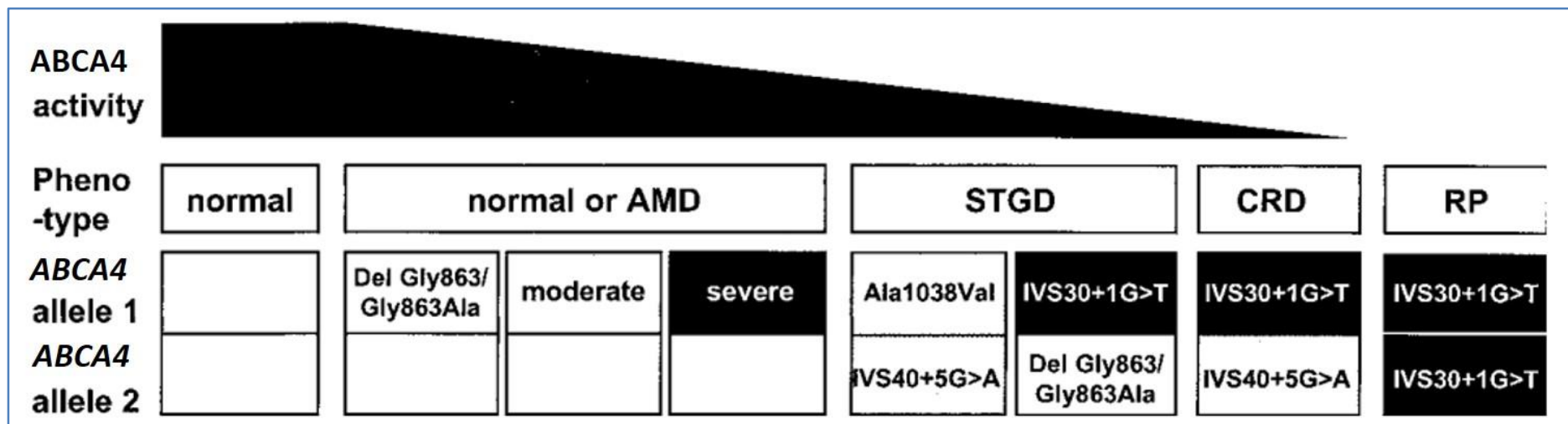


Figure 6.7 Correlation between severity of *ABCA4* mutations and phenotypes. More severe mutations result in a CRD or RP phenotype, whilst milder mutations cause STGD or AMD. Image adapted from Maugeri *et al.* (1999). AMD = Age related macular degeneration, STGD = Stargardt disease, CRD = Cone-rod dystrophy, RP = Retinitis pigmentosa.

6.1.1 Detection of mutations in Stargardt disease with different technologies

Since the initial report of mutations in *ABCA4* causing Stargardt disease, many cohorts of patients with the condition have been analysed to identify pathogenic mutations.

Before NGS was available genetic testing relied on a number of different techniques, as introduced in section 1.7.1 p. 51. Allikmets and colleagues (1997b) used single strand conformation polymorphism (SSCP) analysis, in combination with heteroduplex analysis to achieve a 19.8% detection rate of pathogenic variants in the 36 exons of the *ABCA4* gene known at the time (Allikmets *et al.* 1997b).

Later groups such as Maugeri and colleagues utilised SSCP to identify small mutations, Southern blots to identify large deletions, as well as Sanger sequencing over some exons of the gene, and were able to achieve a detection rate of 62.5% of pathogenic variants (Maugeri *et al.* 1999).

Although SSCP was a useful diagnostic method, it required optimisation to the target region of interest, and the sensitivity suffered as the size of amplicons increased (Weber *et al.* 2005). Later work reported it to have an upper limit of approximately 35% detection rate of high penetrance recessive disease causing variants in *ABCA4* when used on its own (Webster *et al.* 2001).

Shroyer and colleagues later used Sanger sequencing across all the known exons of the gene to detect 80% of pathogenic variants in the disease chromosomes of their small patient cohort (Shroyer *et al.* 2001). This represented a great increase in detection rate, however there is a difference between detection rate and the percentage of patients in whom a secure molecular diagnosis can be achieved. The overall percentage of patients in whom a secure molecular diagnosis could be made was lower than 80%, due to there being many patients with a potentially pathogenic variant identified on just one allele.

These studies were the first of their kind in this field and showed the value of genetic testing in this condition. However nearly all of the studies showed limitations. The early reports by

Allikmets almost certainly had lower detection rates as a result of not analysing all of the exons, which were only reported in full a year later. In addition, it is probable that there would have been some differences in the clinical criteria used to diagnose the phenotypes of patients. It is likely that some patients were diagnosed as having Stargardt disease, when in fact they had a condition caused by mutations in a gene other than *ABCA4*.

6.1.2 Screening of the *ABCA4* gene in STGD patients in the UK

Despite the discovery that mutations in *ABCA4* caused a variety of macular dystrophies, genetic testing on a clinical basis was not available in the UK for many years because Sanger sequencing costs were too high, especially for such a large gene. Very few laboratories offered diagnostic tests, and most patients had a clinical diagnosis without molecular confirmation, leading to confusion for clinicians and patients. This changed when Asper diagnostics developed the arrayed primer extension (APEX) array to screen for *ABCA4* mutations (Jaakson *et al.* 2003). This could detect variants efficiently, however it was only able to detect variants which had previously been reported and which were included in the probe design.

In 2005-2007 Prof Susan Downes (Ophthalmic surgeon) and Prof Andrea Nemeth (Clinical genetics) at the University of Oxford started collecting patients with a clinical diagnosis of STGD. The aim was to determine the clinical utility of genetic testing of *ABCA4* and decide which technology would be the most cost efficient and accurate to introduce into a clinical diagnostic setting.

DNA samples from a series of 50 patients with a macular dystrophy, of whom 34 were considered to have a clinical phenotype most compatible with STGD were analysed using Sanger sequencing of 50 *ABCA4* exons, plus flanking sequence as described in section 6.2 p. 219. Those patients with one or no mutations identified through Sanger sequencing were also analysed using Multiplex ligation probe amplification (MLPA) to test for large deletions. This screening method was able to find at least two potentially pathogenic variants in 27/34 patients

with Stargardt disease, yielding a 79% molecular diagnosis rate, the highest yet reported (Downes *et al.* 2012). This was almost certainly the result of excellent clinical phenotyping, combined with the use of Sanger sequencing, instead of an indirect screening method such as SSCP or APEX arrays. The 27 patients diagnosed with STGD, and in whom two pathogenic variants in the *ABCA4* gene had been identified are listed in **Table 9.1** in appendix section 9.2 p.314.

The remaining 16 patients in the cohort were diagnosed with a MD phenotype other than STGD. In 8/16 patients two potentially pathogenic variants in the *ABCA4* gene were identified. The majority of these patients had a clinical diagnosis of STGD but had progressed with time to a cone-rod dystrophy. These patients are listed in **Table 9.2** in appendix section 9.2 p.314.

In total 35/50 of the original cohort had two likely pathogenic mutations in *ABCA4*. Most had a phenotype of STGD1, but as noted above in some this was an historical diagnosis with a later phenotype of CRD.

The remaining 15 cases (7 patients with a STGD1 phenotype and 8 patients with other MDs) had only a single likely pathogenic variant in the *ABCA4* gene. This cohort of “single pathogenic variant” patients was later expanded as cases were prospectively collected, increasing the size of the cohort to 22 patients. In total this cohort of 22 “single pathogenic variants” included 8 patients who had a MD phenotype other than STGD, and 14 patients with a clinical diagnosis of STGD. These patients are shown in **Tables 6.1** and **6.2** respectively.

The 14 STGD patients listed in **Table 6.2** support an observation reported by many other groups, as previously described, in which a significant minority of STGD patients have only one detectable mutation in *ABCA4*. In these cases, it is highly likely that there is a second mutation in *ABCA4*, since STGD1 is an autosomal recessive disorder and therefore two mutations (in trans) are expected. The question of where the missing mutations in these patients are, and why we have not yet been able to find them, is an important problem. Finding the second mutation is required to be able to provide a secure molecular diagnosis for these patients. This is

necessary to provide accurate genetic counselling, sometimes provide a reliable prognosis and ultimately may help determine future therapies.

The problem is also of interest because STGD is not the only autosomal recessive disease in which two pathogenic variants cannot be found in some cases. If the answer to why the missing variants have not been found so far in *ABCA4* is elucidated, it will likely provide a possible explanation for other recessive conditions. It was this question that initiated the basis of this research project.

In the more heterogeneous MDs in which a single *ABCA4* variant was identified, although it is still possible that the disease is an *ABCA4* retinopathy and that a second *ABCA4* mutation could be present to confirm this. Alternatively, there may be another explanation for the cause of the disease, as will be addressed below.

Table 6.1 Patients diagnosed with a non-Stargardt disease macular dystrophy phenotype in whom only a single potentially pathogenic variant in the *ABCA4* gene has been identified.

Patient ID	Phenotype	Potential pathogenic variants identified		
		Exon/Intron	Nucleotide change	Protein change
245F6	BEM	Exon 33	c.4685 T>C	p.I1562T
381E1	CRD (historical STGD)	Exon 6	c.655 A>T	p.R219X
646B1	BEM	Exon 38	c.5381 C>T	p.A1794D
724B6	PD	Exon 12	c.1726 G>C	p.D576H
749E4	FF-like		c.161 G>A	p.C54Y
803H2	CRD (historical STGD)	Exon 6	c.655 A>T	p.R219X
1049H2	Atypical MD, possibly STGD	Exon 16	c.2496 C>T	p.D832D
1089E2	Unusual FF-like/DD	Exon 14	c.1964 T>G	p.F655C

BEM = Bull's eye maculopathy, CRD = Cone-rod dystrophy, PD = Pattern dystrophy, FF = Fundus flavimaculatus, MD = Macular dystrophy, DD = Dominant drusen.

Table 6.2 Patients diagnosed with a Stargardt disease compatible phenotype in whom only a single potentially pathogenic variant in the *ABCA4* gene have been identified.

Patient ID	Phenotype	Potential pathogenic variants identified		
		Exon/Intron	Nucleotide change	Protein change
189K0	STGD	Exon 23	c.3364 G>A	p.E1122K
475F7	STGD	Exon 44	c.6089 G>A	p.R2030Q
520C7	STGD	Exon 38	c.5461-10 T>C	
531B2	STGD	Exon 5	c.498 delT	p.L167fs
562H2	STGD	Exon 10	c.1253 T>C	p.F418S
628B8	STGD	Exon 30	c.4469 G>A	p.C1490Y
658B6	STGD	Exon 39	c.5537 T>C	p.I1846T
746G4	STGD	Exon 20	c.2966 T>C	p.V989A
781A3	STGD	Exon 42	c.5882 G>A	p.G1961E
788G4	STGD	Exon 15	c.2345 G>A	p.W782X
845F8	STGD	Exon 17	c.2588 G>C	p.G863A
968E3	STGD	Exon 38	c.5461-10 T>C	
1062F7	STGD	Exon 38	c.5461-10 T>C	
1127F8	STGD	Exon 38	c.5461-10 T>C	

6.1.3 Aims of the work presented in this chapter

The aims of the work described in this chapter were as follows:

- 1) To identify a molecular explanation for the cause of the condition in the cohort of patients with a macular dystrophy phenotype other than Stargardt disease.

Macular dystrophies can show a great deal of heterogeneity, so in the cohort of patients with a macular dystrophy phenotype other than STGD, it was considered possible that some patients had a mutation(s) in a gene other than *ABCA4*. This could most easily be tested by using NGS to perform targeted sequencing on a panel of genes that are known to be associated with retinal degeneration, or by performing whole genome sequencing.

- 2) To identify a second pathogenic variant in the *ABCA4* gene in the patients diagnosed with Stargardt disease in whom a single potentially pathogenic variant has so far been identified, or to find another explanation for only one variant being found.

In the cohort of patients with STGD the presence of a single potentially pathogenic variant already identified, combined with excellent clinical phenotyping strongly suggests that the condition of the patients is being caused by mutations in the *ABCA4* gene. A dominant effect of a heterozygous mutation can be ruled out because of the pattern of inheritance in the families and the fact that the carrier parents did not have STGD. The second aim of this chapter is therefore to identify the missing pathogenic variant in the *ABCA4* gene in these patients.

6.1.4 Hypotheses to account for why a second pathogenic variant in *ABCA4* cannot be identified in some patients with STGD

There are a number of possible hypotheses that could account for why a second potentially pathogenic variant in the *ABCA4* gene could not be found in these patients:

1. Mutations in a different genetic locus

Although mutations in *ABCA4* are the most likely explanation, there are several phenocopies of Stargardt disease caused by mutations in other genes. The patients in this cohort may have STGD3 caused by mutations in *ELOVL4* (Zhang *et al.* 2001) or STGD4 caused by mutations in the *PROM1* gene (Yang *et al.* 2008), or they may even have a previously unreported type of Stargardt disease caused by an as yet unknown gene.

Macular dystrophies show a great deal of phenotypic heterogeneity, so the patients may have another type of macular dystrophy with a similar appearance to Stargardt disease caused by variants in another gene. All these possibilities can be tested using next-generation sequencing and a capture design that targets other retinal genes.

2. Cryptic exons caused by intronic mutations

Intronic nucleotides by definition do not encode amino acids as they are spliced out of the transcribed RNA sequence before it becomes mature messenger RNA. However, this process relies on the use of splice site recognition sequences to which the splicing apparatus binds to remove the introns. If a mutation occurs in the intronic DNA it could potentially create such a sequence, causing part of the intronic DNA to become a newly activated cryptic exon, which may have a pathogenic effect, in some cases by creating a frameshift, or by disrupting the translation of a functionally important part of the protein. Such instances have been previously reported to cause other retinal dystrophies (Webb *et al.* 2012).

Although the exons had been Sanger sequenced, the intronic regions had not previously been sequenced in the patients in the cohort studied in these investigations, so this could potentially explain why pathogenic variants had not been identified.

3. Additional exons

The location of exon/intron boundaries in the *ABCA4* gene might not be accurately annotated. In 1998 Gerber and colleagues reported that the *ABCA4* gene contained 50 exons (Gerber *et al.* 1998). However, during the course of the investigations described in this chapter, a paper was published in which intronic variants in *ABCA4* were shown to be pathogenic (Braun *et al.* 2013). The paper reported seven variants as pathogenic, which were located in an alternately transcribed cryptic exon in intron 36, named “exon 36.1” (Braun *et al.* 2013). Braun and colleagues found that mutations in this alternatively transcribed exon could increase the expression of transcripts containing the additional exon, creating a pathogenic effect (Braun *et al.* 2013). The exact mechanism by which this additional exon causes pathogenicity is not reported in their paper, however it would be expected to disrupt a functional domain. The novel pathogenic variants were reported to account for half of the missing pathogenic variants in their cohort. This still left 10% of the total patients with Stargardt disease in their cohort without an explanation for the cause of their condition.

During the course of these investigations a paper was published by Farkas and colleagues in which they performed high depth RNA-seq using mRNA prepared from three normal human retinas (Farkas *et al.* 2013). They reported the discovery of 79,915 novel alternative splicing events, including 29,887 novel exons (Farkas *et al.* 2013). This data shows exons in the *ABCA4* gene that are not annotated in the RefSeq, including exon 36.1 identified by Braun and colleagues, as well as several others.

The data are available online to view as an additional track on UCSC genome browser (<http://oculargenomics.meei.harvard.edu/index.php/ret-trans>). The data include previously unannotated coding regions in the *ABCA4* gene. This could provide a potential explanation for

why pathogenic variants have not yet been identified, as they may be present in a sequence that was previously not known to be coding.

4. Copy number variations, deletions, duplications, or rearrangements

Copy number variations (CNVs) are duplications or deletions of large regions of the genome, which can include whole genes. CNVs can cause disease through a variety of methods such as gene disruption, gene dosage and positional effects (reviewed by Zhang *et al.* 2009).

Copy number variants in the Complement Factor H (*CFH*) gene were suggested to have an association with age related macular degeneration (AMD) by Liu *et al.* (2011), who used a real time-PCR based copy number assay to suggest that fewer copies of the gene were present in patients with AMD. This would support the proposed role of *CFH* in protecting against the disease by reducing chronic local inflammatory and immune-mediated events in the Bruch's membrane which cause drusen biogenesis, and consequently, AMD (Anderson *et al.* 2010).

Whole gene deletions or large in-frame partial deletions on one allele might not be detected by Sanger sequencing, but they can be detected by MLPA (Marcinkowska-Swojak *et al.* 2013). The MLPA performed on this cohort showed normal exons in all of the patients, however this was only performed on known annotated exons, so a large deletion of intronic DNA where an undescribed exon exists may not have been detected.

Large deletions can be also detected by the appearance of homozygosity over a large region of sequence data, which is actually hemizyosity. Duplications and rearrangements can both be difficult to detect in NGS data without long sequencing reads.

5. Tissue mosaicism

The variant may be present only in the tissue of the retina, and not in the lymphocytes which were tested, due to it having arisen as a somatic mutation during development. Identifying such variants would require testing of retinal tissue, which would be far too invasive to be ethically justifiable.

6. miRNA (micro RNA) binding

Micro RNA (miRNA) binds to the transcribed messenger RNA (mRNA), preventing its translation and speeding up deadenylation, and in turn causing the mRNA to break down sooner (Eulalio *et al.* 2009). This can be detected by searching the sequencing data to see if any mutations are creating a change in the sequence that is complementary to known miRNAs. This approach is difficult as the miRNA sequence that binds to the mRNA does not have to match exactly to the mRNA sequence, and it may also span an exon junction, making it difficult to find in a genomic DNA sequence.

7. Repeat sequences

Trinucleotide repeat expansions (TREs) have been shown to cause diseases such as Friedrich's ataxia, where an unstable GAA trinucleotide repeat in the first intron of the *X25* gene causes the majority of cases (Campuzano *et al.* 1996), or Huntington's disease in which many cases are caused by a CAG repeat sequence. The *ABCA4* gene has several GAA and CAG trinucleotide repeat sequences in the gene where a potential repeat expansion could occur. If such an expansion has occurred and it is very large it may be detected by agarose gel electrophoresis of Long Range PCR products.

There are other types of repeat such as Alu repeats, and long interspersed nuclear elements (LINE) present in the genome. These repeats can cause rearrangements of the DNA to occur. They are usually small and cannot be detected on agarose gels. To detect these using NGS, long reads and paired end reads are both useful to allow mapping of the repeat regions.

8. Coding changes in the *ABCA4* gene having unknown pathogenic effects

The Sanger sequencing already performed on these patients should have identified any coding changes in the fifty known exons of the *ABCA4* gene, which would make it unlikely that the second pathogenic variant is causing a coding change in a known exon of *ABCA4*, although

bioinformatic prediction programs are not perfect and may incorrectly identify variants as benign rather than pathogenic. Alternatively, the variant could be in a coding region, but causing a change to an exonic splice enhancer or silencer.

The splicing of introns from pre-mRNA to form mature mRNA requires several components. The splice sites at the 5' and 3' end of each exon have already been discussed, but there are also exonic splice enhancers (ESEs) and exonic splice silencers (ESSs), six-base sequences located near the centre of an exon to which the spliceosome complex binds. These hexamers can either assist or prevent splicing from occurring. Mutations in these sequences have been reported to cause disease through exon skipping in the same way that mutations to splice sites do (Stella *et al.* 2001). Potentially pathogenic mutations in ESEs and ESSs are hard to detect as there have been hundreds of hexamer sequences identified so far that can act as ESEs and ESSs. Examples of such changes that have been reported include the disruption of an ESE in the Survival of motor neuron 2 (*SMN2*) gene causing proximal spinal muscular atrophy in the absence of another gene *SMN2* (Cartegni and Krainer 2002), and a c.30C>G (p.Gly10Gly) substitution in exon 1 of the phenylalanine hydroxylase (*PAH*) gene which creates an ESS sequence that causes skipping of the exon, ultimately causing phenylketonuria (Dobrowolski *et al.* 2010).

9. Mutations in promoter, enhancer, or silencer regions

Genes are regulated by promoters, enhancers and silencers, which play important roles in transcription. Core promoters are sequences upstream of the transcription start site to which the RNA polymerase and a variety of transcription factors bind, allowing transcription to initiate. Proximal promoters lie further upstream and contain sequences to which other transcription factors can bind to regulate the level at which transcription occurs. Enhancer and silencer sequencers can be located further still from the gene, and regulate the activity of proximal promoter sequences. Mutations to any of these sequences can modify how this occurs, reducing

or increasing the level of transcription of a gene, causing a disease phenotype as a result (reviewed in Maston *et al.* 2006).

Mutations to promoter sequences are a known mechanism of disease pathogenicity. Reported cases include a c.1-133C>G mutation in the promoter of the platelet glycoprotein Ib, beta polypeptide (*GP1BB*) gene which causes a loss of function and results in Bernard-Soulier syndrome (Ludlow *et al.* 1996), and a c.1-458C>T mutation in the promoter sequence of the Gap junction protein, 32-kD (*GJB1*, formerly known as nerve-specific connexin32) gene which causes dominant Charcot-Marie tooth syndrome (Ionasescu *et al.* 1996).

Most regulatory mutations that have so far been identified and reported in the Human Gene Mutation Database (www.hgmd.org) are located between +50 and -500 of the transcription start site (Stenson *et al.* 2003), so screening these regions in the *ABCA4* gene for potential mutations to these sequences could provide another possible explanation for the cause of disease.

10. Epigenetic effects

It may be possible that there are no variants in the actual genomic DNA sequence, and that instead methylation of the DNA on one allele prevents it from being expressed. Methylation of DNA can be determined using the bisulphite sequencing method (Frommer *et al.* 1992).

6.1.5 Investigations that were performed in the work described in this chapter

The list of possible explanations to account for why a second pathogenic mutation could not be found in the *ABCA4* gene in some cases is lengthy, and investigating every one of these hypotheses in the time available was not possible. Instead several hypotheses were chosen to be tested based on their predicted probability of being correct, and on the ease with which they could be tested.

6.1.5.1 Searching other genetic loci using targeted NGS

The cohort of patients in this chapter included patients with Stargardt disease, and also patients with a range of other macular dystrophy phenotypes. It was considered probable that in at least some cases the cause of the patients' condition was a mutation in a gene other than *ABCA4* due to the heterogeneity of macular dystrophies. The ability to economically sequence multiple genes is one of the main advantages of NGS, so the first aim of the work described in this chapter was to perform targeted capture and sequencing of 74 genes associated with retinal disease, and to filter and analyse the data generated to identify pathogenic mutations in the 73 genes other than *ABCA4*.

6.1.5.2 Identifying cryptic exons in targeted sequencing data

Although in some patients the cause may have been another gene, this was extremely unlikely to be the case in the cohort of patients which had been diagnosed clinically with Stargardt disease, especially as in each of these cases a single potentially pathogenic variant in the *ABCA4* gene had already been identified. Data from the intronic regions of the *ABCA4* gene generated from targeted sequencing was therefore analysed using five splice prediction programs to determine whether intronic variants were activating cryptic exons that could have a potentially pathogenic effect.

6.1.5.3 Capturing DNA using Long Range PCR for sequencing

One possible explanation for the failure to detect a second pathogenic variant in the *ABCA4* gene is that the condition was caused by a large intronic deletion or complex rearrangement (such as an inversion) which would not be detected by previous screening methods. Most NGS methods use short reads which might not be long enough to cover a large sequence deletion. The bioinformatic processing of the data usually includes aligning the reads to a reference sequence, which can determine if a deletion has occurred by the lack of reads covering that

region. However, this methodology can result in reads which contain the breakpoint in a rearrangement being discarded as falsely erroneous.

It was therefore decided to attempt next-generation sequencing of the patients using Long Range PCR (LR-PCR) as a sequence capture and enrichment method. By using paired end reads with a long insert for sequencing it should be possible to more accurately identify any possible large rearrangements or deletions that might have been missed by other methods.

LR-PCR also has the advantage of only requiring the small regions to which the primers bind to accurately match the target design. All sequence amplified will then be sequenced regardless of whether it differs from the reference sequence.

6.1.5.4 Whole genome sequencing

Whole genome sequencing (WGS) removes any bias that might be introduced to sequencing by a capture and enrichment method. WGS was therefore used on samples from the cohort of patients with a single potentially pathogenic variant in *ABCA4*. The data were analysed firstly to determine whether the previously reported pathogenic variants were detected by this method, using these as positive controls to test the technology. Secondly, the data were analysed to determine whether any new variants were identified which could potentially be pathogenic. An analysis was also made of the data to search for mutations in other genetic loci that might cause macular dystrophies using Ingenuity Variant Analysis software (see section 2.1.5 p. 73).

6.1.5.5 Expression of *ABCA4* transcripts

Mutations in genomic DNA that have an effect on phenotype cause changes to the RNA they transcribe. It was therefore hypothesised that the missing pathogenic mutations in *ABCA4* could be identified by extracting mRNA from the patients, reverse transcribing it into cDNA, and then

identifying variation either using agarose gel electrophoresis or sequencing, as had previously been performed for the early onset epilepsy patients described in Chapter 5.

For the early onset epilepsy patients described in Chapter 5 it was possible to investigate the expression of the genes being studied by extracting mRNA from blood, as the genes were expressed in leukocytes. The *ABCA4* gene however is only expressed at significant levels in a small number of tissues other than the retina, and is expressed at very low levels in leukocytes. A paper published by Maugeri and colleagues described a technique in which *ABCA4* was successfully amplified using nested PCR of cDNA reverse transcribed from mRNA extracted from lymphoblastoid cells which had been immortalised with the Epstein-Barr virus (Maugeri *et al.* 1999). A later paper by Beit-Ya'akov and colleagues used the same method with different primers to amplify a different region of the gene (Beit-Ya'acov *et al.* 2007). It was decided to attempt to use the same method described in these papers for cDNA of *ABCA4* extracted from lymphoblastoid cells using this method.

Once successfully performed the process could be repeated, and targeted to areas where the single known pathogenic variants reside, especially in the patients with stop codon variants. If a known pathogenic variant is detectable by sequencing it shows that it is not causing the mRNA to be degraded by nonsense-mediated decay. If a known pathogenic variant is not detected, the reverse would be indicated to be true.

If no cDNA is amplified it might indicate that on both alleles there are pathogenic mutations which both cause nonsense-mediated decay, and as a result no protein at all is being produced by the gene. However, it would be unlikely that such severe *ABCA4* mutations would cause an STGD phenotype.

6.1.5.6 Tests performed on each patient

Table 6.3 shows which of these experiments were performed on each of the patients listed in **Table 6.1** (p.204) and **Table 6.2** (p.205).

Table 6.3 Experiments performed on the patients with macular dystrophy phenotypes.

Patient ID	Phenotype	Experiments performed		
		Targeted sequencing	Whole genome sequencing	Long Range PCR amplification and sequencing
245F6	BEM	Yes		
381E1	CRD (historical STGD)	Yes	Yes	
646B1	BEM			
724B6	PD			Yes
749E4	FF-like		Yes	Yes
803H2	CRD (historical STGD)		Yes	
1049H2	Atypical MD		Yes	Yes
1089E2	Unusual FF-like/DD			Yes
189K0	STGD		Yes	Yes
475F7	STGD	Yes		
520C7	STGD		Yes	Yes
531B2	STGD	Yes		Yes
562H2	STGD			
628B8	STGD			
658B6	STGD			Yes
746G4	STGD		Yes	Yes
781A3	STGD		Yes	Yes
788G4	STGD		Yes	Yes
845F8	STGD			
968E3	STGD			Yes
1062F7	STGD			Yes
1127F8	STGD		Yes	Yes

6.2 Methods and materials

The experiments described in this chapter utilised four different experimental techniques:

- Targeted sequencing using three different combinations of sequence capture and next-generation sequencing platform (as described in Chapter 3).
- Whole genome sequencing (performed by Complete Genomics).
- Long Range PCR generated amplifications of the *ABCA4* gene, sequenced on the Illumina HiSeq 2500.
- Extraction of mRNA from cell lines derived from leukocytes, reverse transcription into cDNA, nested PCR, and Sanger sequencing.

These experiments were performed on nucleic acid samples taken from two cohorts of patients:

1. Eight patients who had been diagnosed with a macular dystrophy phenotype other than Stargardt disease through ophthalmological examination, and in whom only a single pathogenic variant in the *ABCA4* gene had been identified. These patients are listed in **Table 6.1** section 6.1.2 p. 204.
2. Fourteen patients who had been diagnosed with Stargardt disease through ophthalmological examination, and in whom only a single pathogenic variant in the *ABCA4* gene had been identified. These patients are listed in **Table 6.2** section 6.1.2 p. 204.

Each of these patients is referred to in this chapter by an anonymous patient ID, for example patient “245F6”.

6.2.1 Targeted next-generation sequencing

Targeted next-generation sequencing was performed on DNA samples from four patients: 245F6, 381E1, 475F7, and 531B2. The sequencing was performed on each patient three times using three different combinations of sequence capture and NGS technology, as described in section 3.2 p.98.

These three combinations of technology were:

- NimbleGen sequence capture array and Roche 454 GS FLX.
- RainDance and Roche 454 GS FLX.
- SureSelect and Illumina HiSeq 2000.

Data generated from the sequencing was filtered and analysed using the improved bioinformatic pipeline described in section 4.3.2 p.143 to search for pathogenic variants in 74 retinal genes.

In addition, the entire sequence of the *ABCA4* gene including all introns and exons was captured, sequenced, and the data analysed, to search for mutations in introns that might create a cryptic exon by activating a cryptic splice site. A list of such potential variants was compiled by filtering out exonic variants. The remaining intronic variants were analysed by Alamut mutation prediction software using four different splice prediction programs:

- SpliceSiteFinder-like (<http://www.genet.sickkids.on.ca/~ali/splicesitefinder.html>).
- MaxEntScan (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html).
- NNSPLICE (http://www.fruitfly.org/seq_tools/splice.html).
- GeneSplicer (http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml).
- Human Splicing Finder (<http://www.umd.be/HSF/>).

Those that showed more than a 10% change in score on each program were annotated for their frequency in dbSNP. Variants with less than 20% frequency were amplified with PCR and Sanger sequenced to confirm their authenticity as described in sections 2.2.6 p.79 and 2.2.8.1 p.87 respectively.

6.2.2 Whole genome sequencing using ligation based sequencing and DNA nanoballs

DNA extraction from whole blood samples of patients was performed by the Oxford Regional Genetics Laboratory (ORGL) using Qiagen Autopure LS and EZ1 systems supplied by Qiagen (UK).

DNA samples were diluted to a concentration of 75 to 300ng/μl and a volume of 50 to 200μl in 1x TE pH 8.0. Concentrations were determined by Qubit fluorometer (see section 2.2.3.2 p.77). A small volume of each sample was used for a 1:1,200 dilution which was run on a 0.8% agarose gel to test for DNA quality. Samples that passed quality control showed one clear strong band above a 20kbp marker ladder upon ultraviolet imaging using a transilluminator.

Samples were shipped to Complete Genomics who performed all stages of library preparation and sequencing. Data was returned both as variant called files (VCFs) and in Complete Genomics own master variant (MasterVar) file formats. Full details of the assembly pipeline and variant calling thresholds used by Complete Genomics can be found in a publication by Carnevali *et al.* (2012).

For the data generated in this project, a pass threshold of 20 was applied for homozygous variants. Variants with scores below 20 were considered LOW quality, and those above PASS quality. For heterozygous variants the threshold used was 40.

MasterVar files were uploaded to Ingenuity Variant Analysis servers to perform analysis for variants using the keywords “Stargardt disease” and “macular dystrophy”.

Further analysis was manually performed on the raw data using the annotations of the retinal transcriptome published by Farkas and colleagues (available at <http://oculargenomics.meei.harvard.edu/index.php/ret-trans>), as well as data from the paper by Braun and colleagues to identify variants that were close to the novel exons identified by these papers (Braun *et al.* 2013, Farkas *et al.* 2013).

6.2.3 Long Range PCR generated amplifications of the *ABCA4* gene, sequenced on the Illumina HiSeq 2500

Long range PCR was performed as described in section 2.2.6.3 (p. 83) using the primers in **Table 2.3** to generate amplicons covering the entire *ABCA4* gene, as shown in **Figure 6.8**.

6.2.3.1 SequalPrep Normalization Plates

A SequalPrep Normalization Plate Kit containing SequalPrep normalization plates, SequalPrep Normalization Binding Buffer, SequalPrep Normalization Wash Buffer, and SequalPrep Normalization Elution Buffer) comprised of 10mM Tris-HCl, pH 8.5, all provided by Invitrogen (UK), was used for the purification and equimolar pooling of amplicons produced by LR-PCR.

The SequalPrep normalization plates were manufactured with a positive surface charge. When PCR products were pipetted into the wells and mixed with a low pH buffer (the SequalPrep Normalization Binding Buffer), the positive surface of the wells attracted the negatively charged backbone of the DNA, binding it to the surface of the wells.

The supernatant was removed, and a wash buffer (SequalPrep Normalization Wash Buffer) was added to the wells, then pipetted up and down, before also being removed. An elution buffer (SequalPrep Normalization Elution Buffer, 10mM Tris-HCl, pH 8.5) was then added to the wells, with the higher pH causing the purified DNA bound to the sides of the wells to be eluted. By using the same volumes of buffer and DNA, covering the same surface area of each well, it was possible to elute equimolar concentrations of DNA from each well which were then pooled.

A comparison of the technology to other normalisation methods for PCR amplicons was performed by Harris and colleagues (Harris *et al.* 2010).

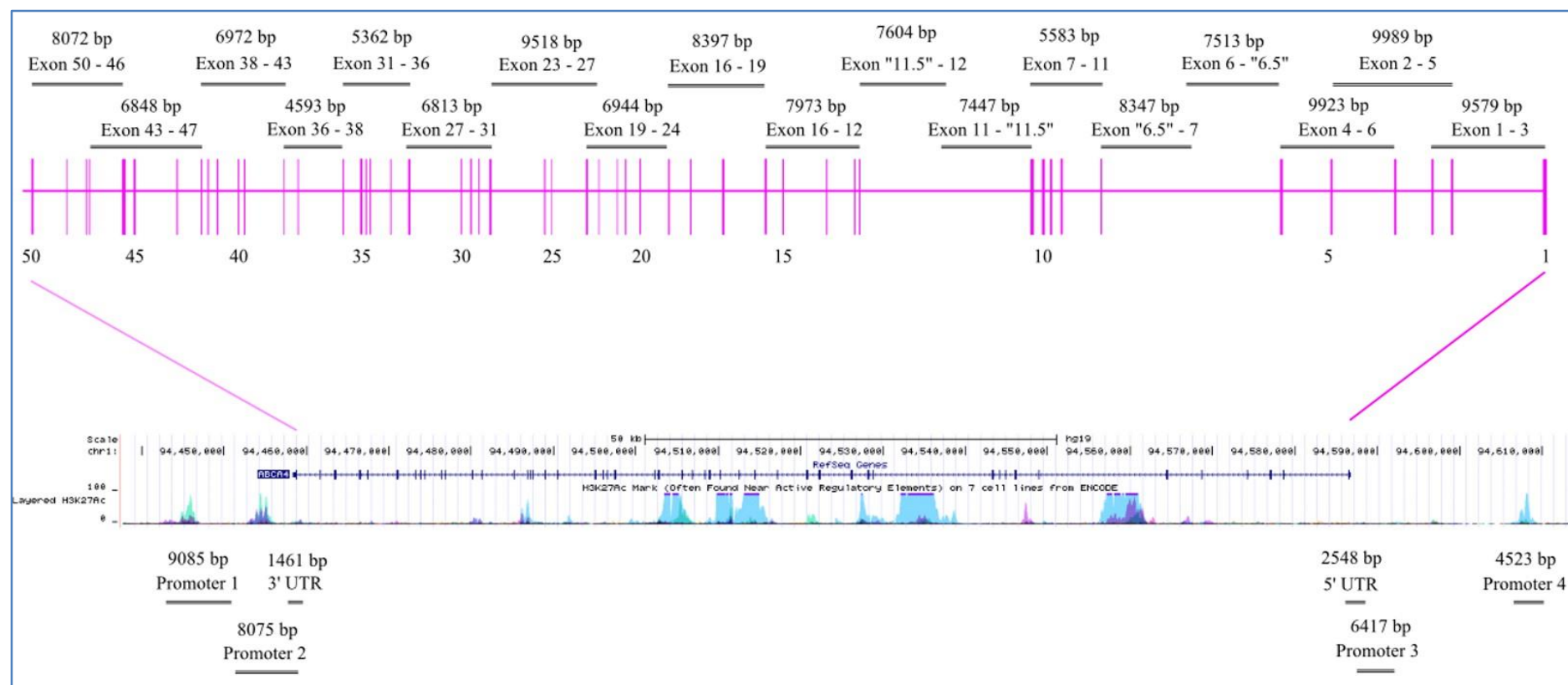


Figure 6.8 Amplicons produced by the LR-PCR amplifications. Schematic showing the positions of the Long Range-PCR amplicons generated by the primer pairs listed in **Table 2.3**, with sizes. The amplifications were designed to cover all the introns of the *ABCA4* gene and promoter regions. Where possible primers were located inside exons and with overlap of amplicons generated, however this was not always possible due to the maximum size of approximately 10kbp amplicons that could efficiently be generated by LR-PCR.

6.2.3.2 Illumina sequencing and bioinformatic processing

Sequencing of the pooled amplicons was performed by the High Throughput Sequencing group at the Wellcome Trust Centre for Human Genetics using an Illumina HiSeq2500 with a standard library preparation for PCR amplicons. Bioinformatic assembly of a continuous sequence from the raw data was performed by Leo Goodstadt (University of Oxford) using Stampy (section 2.1.5 p. 73).

6.2.4 Reverse transcription and amplification of *ABCA4* cDNA

The cDNA experiments described in this chapter were comprised of four steps; generation of lymphoblastoid cell lines from patient blood samples, extraction of mRNA from the cell lines, reverse transcription into cDNA, and finally PCR and nested PCR to amplify targeted sequences of *ABCA4* cDNA.

6.2.4.1 Generation of cell lines and mRNA extraction

Whole blood was extracted by venepuncture from several of the patients in the cohort with a maculopathy by staff in the Oxford Eye Hospital. The lymphocytes in the whole blood were extracted and transformed with the Epstein-Barr virus to grow cell lines (performed by Dr Susan Butler at the Weatherall Institute of Molecular Medicine, University of Oxford). Cell lines were grown, and then pelleted and suspended in TRIzol (supplied by Life Technologies (UK), and performed by Dr Susan Butler). An acid guanidium thiocyanate-phenol-chloroform extraction was carried out on the cells suspended in TRIzol, as described below.

The homogenised samples were incubated at room temperature for five minutes to allow complete dissociation of the nucleoprotein complex. 0.2ml of chloroform was added per 1ml of TRIzol used and the tubes were shaken vigorously by hand for fifteen seconds. The samples were incubated for 3 minutes at room temperature, then centrifuged at 15,000 x *g* for 15 minutes. This separated the solution into three phases; an uppermost, clear, aqueous phase containing

RNA; a cloudy, interphase containing DNA; and a red, lower, organic phase containing lipids and proteins. The aqueous phase was removed by tilting the tube at an angle and carefully pipetting it away without disturbing the interphase layer.

The RNA was then precipitated by adding 0.5ml of 100% isopropanol per 1ml of TRIzol used for homogenisation, incubated at room temperature for 10 minutes, and centrifuged at 12,000 x *g* for 10 minutes at 4°C. The supernatant was pipetted away and discarded. The pellet was washed by adding 1ml of 75% ethanol, vortexing, centrifuging at 7500 x *g* for 5 minutes, and removing the supernatant.

The RNA pellet was left to air dry for approximately 10 minutes and then resuspended in 40µl of RNase-free water. The solution was heated at 60°C for 15 minutes and then chilled on ice. RNA quality was determined by nanodrop spectrophotometry (see section 2.2.3.1 p.77). RNA was stored until needed at -20°C.

6.2.4.2 Reverse transcription of cDNA

Reverse transcription of cDNA from mRNA was performed using a QuantiTect reverse transcription kit from Qiagen (UK). Genomic DNA was removed by mixing 2µl of gDNA wipeout buffer with up to 1µg of RNA, made up to a total volume of 14µl with RNase free water and incubating the mixture for 2 minutes at 42°C. The reverse transcription reaction was made by mixing 1µl of Quantiscript Reverse Transcriptase with 4µl of Quantiscript RT Buffer 5x, and 1µl of RT primer mix. This mixture was then incubated at 42°C for 15 minutes, followed by inactivating the enzyme by heating at 95°C for 3 minutes, and allowing the reverse transcribed cDNA to cool on ice. The reverse transcribed cDNA was stored at 4°C until needed.

6.2.4.3 Amplification by PCR and nested PCR

The cDNA was then amplified by PCR as described in section 2.2.6 p.79 using primers that covered several regions of the *ABCA4* gene listed in **Table 6.4**. The products of these amplifications were amplified again in a nested PCR using primers that hybridised to targets located inside the amplified region (also listed in **Table 6.4**). These nested PCR products were separated by agarose gel electrophoresis (described in section 2.2.4 p.78).

Table 6.4 *ABCA4* cDNA primers.

Primer name	Primer sequence (5' -> 3')
737f	AACGTCAACCCCCGACAC
738r	TCCTGTCCGTCAGGTCTTG
351f	CTACCTTTGTGTTTTGGCTCTG
257r	ACTCTGGCAGCATGGTGAG
2359f	TGTCCATGCAGATGATGCTC
2360r	GGAGTCGTGTATTCCTTCTG
2070f	TGCGTGCTATGGCTTACTCGC
2071r	TGGATCCTCCGTTTCCTCTG
<i>ABCA4</i> cDNA E 38/39	TGAACCTGAGCAGCGTCC
<i>ABCA4</i> cDNA E 34/35	ACTGAAGACAACATTAAGGTGTGG
<i>ABCA4</i> cDNA E 37/38	ACCGCCCATCCATACAGC
<i>ABCA4</i> cDNA E 35/36	CTCTCAGAGATTACAGTGCTGACC
<i>ABCA4</i> 9/8	GGATGTTGTTCTTCTGTCATAAGAA
<i>ABCA4</i> 4/3	CAAACCTATAACAACCTCCATCTTGGC
<i>ABCA4</i> 8/7	GCCGATGGATAAACTCTTGAAT
<i>ABCA4</i> 5/4	GAGAGAATTGCAGGAAGAGGA
<i>ABCA4</i> 18/17	TCTGGTTGAACACCCTTCAC
<i>ABCA4</i> 13/12	GACAGGTATTGGGATTCTGGTC
<i>ABCA4</i> 17/16	CATAGTCTCCTGGAAACACCTG
<i>ABCA4</i> 14/13	GACGATTCTTTCATGATCATCC
<i>ABCA4</i> 43/42	AGGAGGCCAAAGCACTCTC
<i>ABCA4</i> 37/36	GGACATCATGAATTATTCCGTG
<i>ABCA4</i> 41/40	CAATCCATTGGGAGAGGAAGA
<i>ABCA4</i> 38/37	GTATGGATGGGCGGTCATT

6.3 Results

6.3.1 Targeted sequencing of other retinal genes

Targeted sequencing of 74 genes (listed in **Table 2.4** p.90) known to be associated with retinal disease was performed as described in section 3.2 p.98, on four patients in whom only one potentially pathogenic variant in the *ABCA4* gene had been identified through Sanger sequencing or MLPA. These were patients 245F6 and 381E1, who had phenotypes of BEM and CRD respectively, and patients 475F7 and 531B2, who were both diagnosed with STGD.

The data were analysed using the bioinformatic filtering pipeline described in section 4.3.2 **Figure 4.12** (p.143) to attempt to identify pathogenic variants in other retinal genes that could be the cause of the condition in these patients.

6.3.1.1 Identification of a variant in a different genetic locus causing a macular dystrophy in a patient with a single ABCA4 variant

Patient 245F6 was a 43-year-old woman who presented with reduced vision in her right eye over a period of 3 months with a positive family history of macular dystrophy. The proband's sister had been diagnosed as having Stargardt disease aged 17 years old. Fundoscopy of the proband revealed a Bull's eye maculopathy with visual acuities of 6/60 OD, and 6/7.5 OS. Autofluorescence imaging of both eyes is shown in **Figure 6.9 A**.

Electrophysiology testing in the proband showed a significant amplitude reduction in the 'b' wave of the rod electroretinogram, with a mild increase in implicit time of the 'a' and 'b' wave, but was otherwise within normal limits. Her affected sister was re-examined at the age of 45 years, and showed visual acuities of 6/60 right and left with marked central atrophy. Autofluorescence imaging shown in **Figure 6.9 B** showed central dense loss of signal consistent with atrophy with an annulus of increased signal, often seen in inherited retinal dystrophies. Electrophysiology testing was not available. Both parents were healthy and in their 70's with no known visual impairment. The father was examined as part of the family survey and had a normal fundus appearance on colour imaging (**Figure 6.9 C left**). However, autofluorescence imaging revealed

central patchy reduction in AF signal consistent with atrophy surrounded by a fringe of increased signal (**Figure 6.9 C right**).

The patient was found to have a single mutation, c.4685T>C (p.Ile1562Thr) in *ABCA4*, resulting which was of unknown significance, but was considered to be pathogenic by the NHS. None of the other *ABCA4* mutations identified by Sanger sequencing were described as being likely to be pathogenic.

Segregation analysis of this single *ABCA4* variant identified in the proband revealed that the father also carried the variant, but the proband's affected sibling did not. It was therefore concluded that this variant was unlikely to be the cause of the retinal disease in the family.

Analysis of the data from the targeted resequencing of 74 genes associated with retinal dystrophies were performed as described in section 3.2 p.98. Interestingly, a heterozygous substitution, c.121C>T (p.Arg41Trp), was found in the cone-rod homeobox (*CRX*) gene. Sanger sequencing confirmed the variant was present in all three affected family members (**Figure 6.9 D, and Figure 6.10**).

This amino acid change is in a conserved region of the protein, which has previously been reported to be a dominant pathogenic mutation that causes cone-rod dystrophy (CRD, Swain *et al.* 1997, Mitton *et al.* 2000). The phenotype was consistent with a mutation in the *CRX* gene, as intra familial phenotypic variation has been previously reported in association with *CRX* (Kitiratschky *et al.* 2008). The family were informed of the diagnosis and counselled that the risk of inheritance for offspring of the affected siblings was 50% in contrast to the low risk from *ABCA4* mutations.

Patient 245F6 was the only patient in whom a candidate pathogenic mutation in a gene other than *ABCA4* was identified that could potentially cause the condition of one of the patients in the cohort.

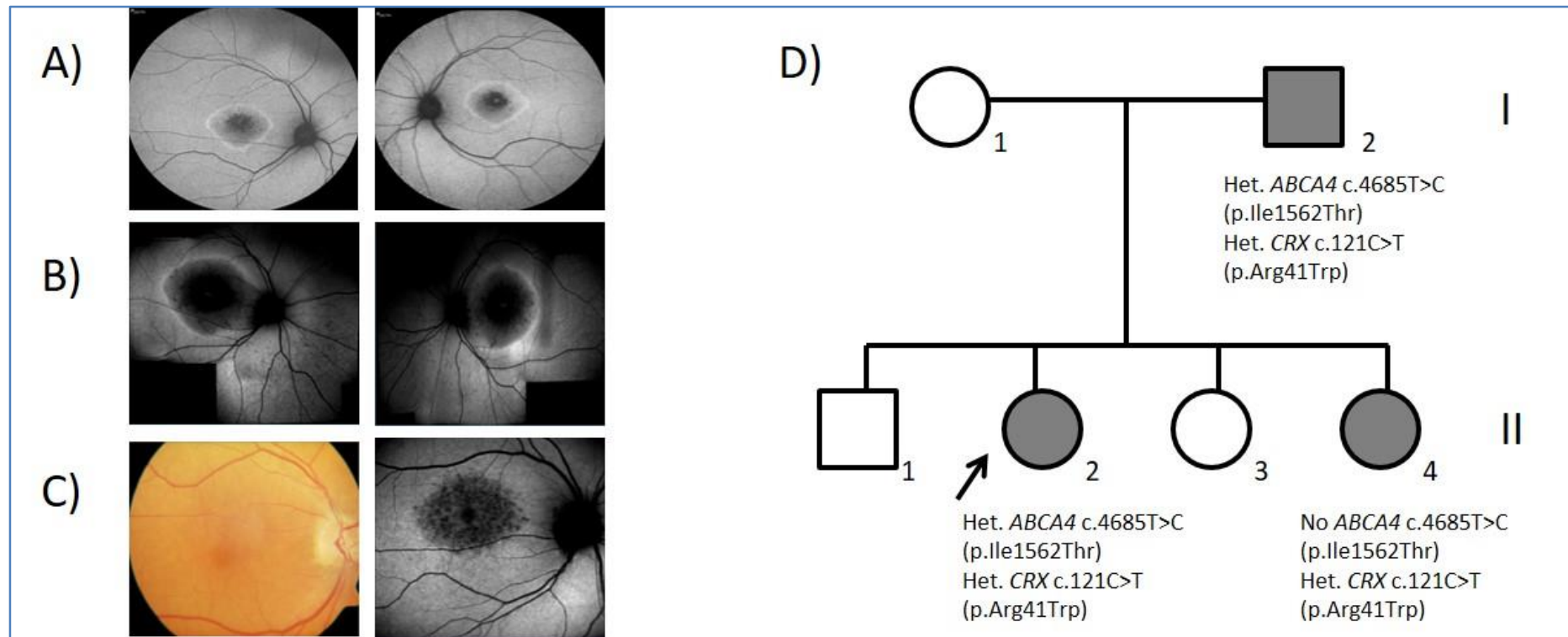


Figure 6.9 Retinal imaging and pedigree of the family of patient 245F6. **A)** Fundus autofluorescence (AF) imaging of the proband (II₂) showing an increased signal at the macula with a central atrophy. **B)** AF imaging of the more severely affected sibling (II₄) shows central dense loss of signal consistent with atrophy with an annulus of increased signal, often seen in inherited retinal dystrophies. **C)** The father (I₂), Colour image of the right eye which appeared normal, but autofluorescence imaging revealed central patchy reduction in AF signal consistent with atrophy surrounded by a fringe of increased signal. **D)** Pedigree of the family. The black arrow indicates the proband. The shaded boxes and circles indicate affected individuals. Het = Heterozygous

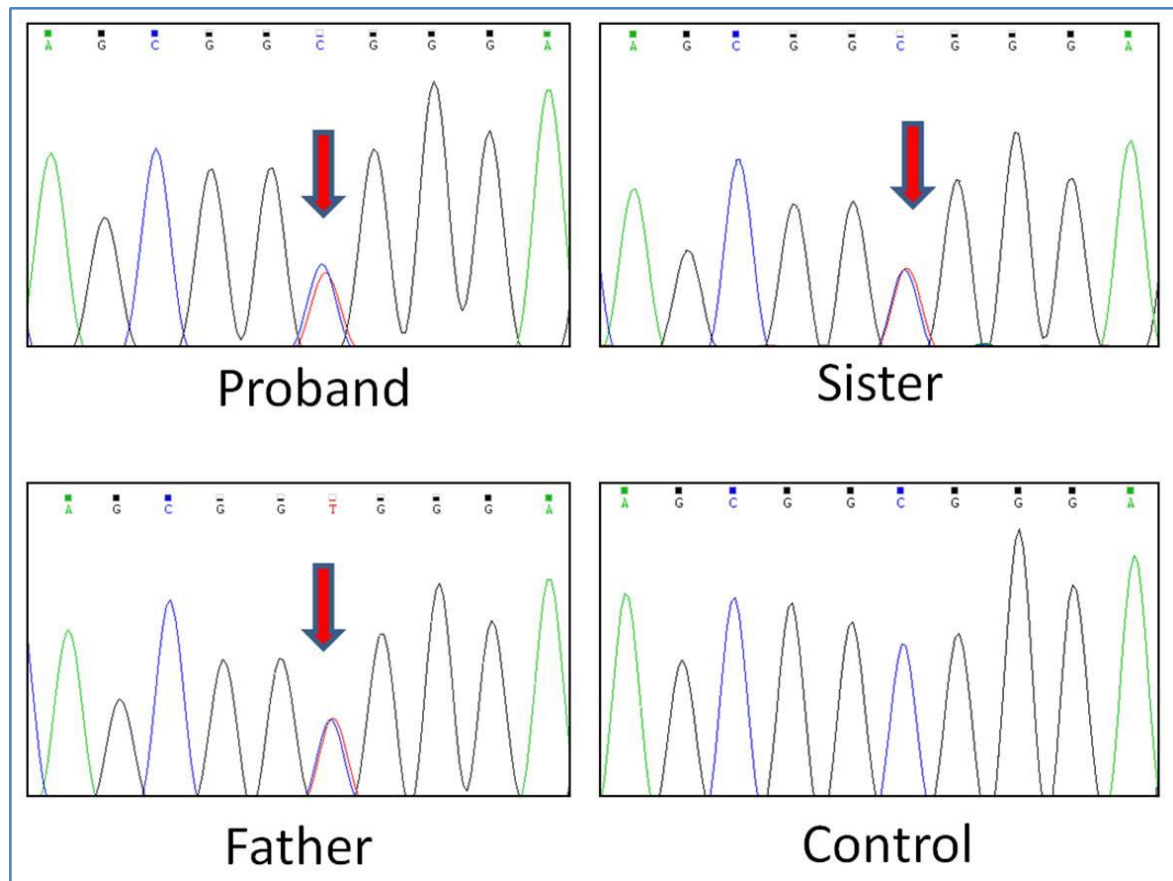


Figure 6.10 Cosegregation of the c.121C>T (p.Arg41Trp) *CRX* variant in the family of patient 245F6. Electropherograms from Sanger sequencing show that the c.121C>T (p.Arg41Trp) *CRX* variant is present in each of the affected family members, and absent in a control.

6.3.2 Re-analysis of targeted sequencing data to search for variants in cryptic exon in *ABCA4* identified by Braun and colleagues

As described in section 3.2 p 98, at the time of the targeted capture of exons of 74 genes, the entire genomic region of *ABCA4* was also captured. The data produced from this *ABCA4* capture was analysed in three patients to determine whether any of the seven pathogenic variants reported by Braun and colleagues in 2013 (listed in **Table 6.5**) were present in any of the patients sequenced in the maculopathy cohort. This analysis was performed on the data from all three combinations of sequence capture and next-generation sequencing platform. Variant 1 in the Braun paper was discovered to be present in patient 487A3 in the data produced by the Raindance and 454 combination, however this variant was not identified in this patient in the data from the other two combinations of NGS and sequence capture technologies.

RainDance uses PCR-based amplification to capture and enrich target DNA, whereas the other two capture methods both rely on hybridisation of probes to target sequence. Hybridisation based capture methods have been shown to have a bias towards reference sequence, as these sequences will match the probes exactly and so bind preferentially compared to mutated sequence (Asan *et al.* 2011). If this is the reason that the RainDance capture method detected the variant from the Braun paper and the hybridisation methods did not, then it is possible that the missing pathogenic variants have not been identified due to the failure of the capture methods used so far to accurately capture sequence in the intronic regions where mutations may be creating cryptic exons that cause pathogenicity.

To further test whether this was the case a capture design was created to cover the entire region of the *ABCA4* gene, including all the intergenic sequence leading up to the neighbouring genes using Agilent SureDesign software for SureSelect capture. This design showed significant gaps in the predicted coverage, due to repeat regions in those part of the genome for which probes could not be designed to cover, including two large gaps in the intergenic sequence on either ends of the gene, as shown in **Figure 6.11**.

Table 6.5 The seven pathogenic intronic or synonymous variants identified by Braun and colleagues (Braun *et al.* 2013).

Variant	Haplotype	Position	Genomic location	ABCA4 (NM_000350.2)	Primary cohort (n=28)	Validation cohort (n = 48)	Controls (n= 300)	1000 genomes (n= 1092)
V1	H1	Exon 36.1-3 G>A	chr1:94,484,001	c.5196+1137G>A	4	4	0	0
V2	-	Exon 36.1+4 C>A	chr1:94,483,922	c.5196+1216C>A	1	0	0	0
V3	-	IVS36+1056 A>G	chr1:94,484,082	c.5196+1056A>G	1	0	0	0
V4	-	Exon 30.1 position 110 G>A	chr1:94,493,000	c.4539+2001G>A	1	0	0	0
V5	H2	Exon 30.1 position 138 C>T	chr1:94,492,973	c.4539+2028C>T	3	4	1	0
V6	-	Val2114Val GTG>GTA	chr1:94,466,602	c.6342G>A	1	2	0	0
V7	H3	IVS33+3 A>G	chr1:94,487,399	c.4773+3A>G	3	0	0	0

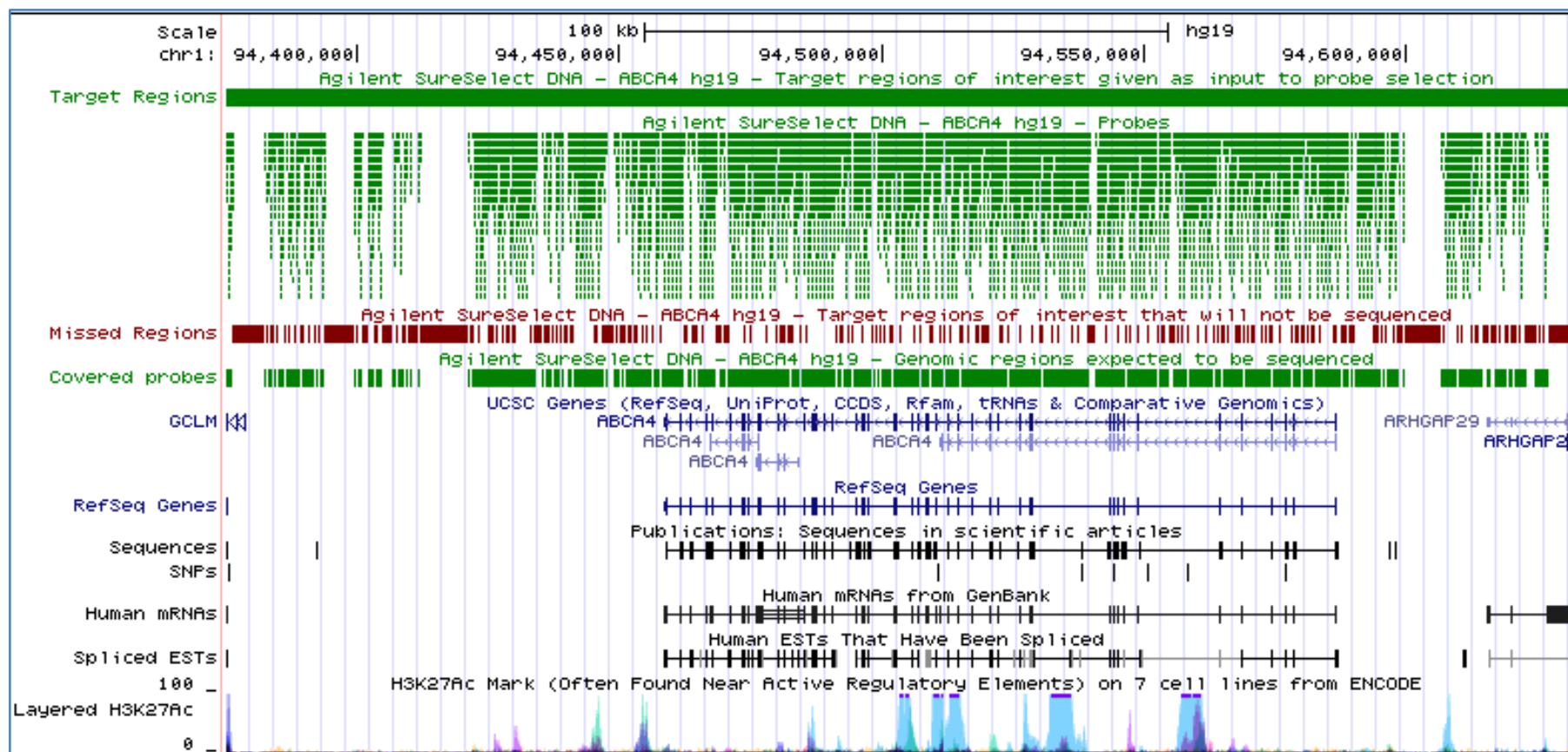


Figure 6.11 Screenshot of the UCSC Genome Browser displaying the genomic region between the *GCLM* and *ARHGAP29* genes. This region includes the *ABCA4* gene in the middle. The top track shows the region which has been input as a target for SureDesign software to create probes for a SureSelect capture. The second track from top shows regions which are missed due to the inability of the software to design probes that will cover these regions, usually due to repeat sequences. The third track from top shows the regions that would be expected to be sequenced by the probes. The next five tracks show the position of the *ABCA4* gene according to different databases, and the bottom track shows H3K27AC marks identified by the ENCODE project, which could potentially signify that there are regulatory elements at these positions. The design displayed here shows gaps in the genomic regions expected to be sequenced, which could prevent variants from being detected.

6.3.3 Long-Range PCR amplification of the *ABCA4* genomic region

The failure of the SureSelect design to be able to capture all the regions in the target could potentially cause pathogenic variants to be missed from sequencing results. This problem could be further exacerbated if the reference sequence itself is not accurate.

To overcome this problem it was decided to amplify the entire genomic region of the *ABCA4* gene through a series of Long Range PCRs, to pool the products, and then sequence them on the Illumina HiSeq 2500, as described in section 6.2.3 p.222. There are two major advantages of this sequence capture and enrichment method.

1. Only the sequence of interest will be amplified, which should lead to a higher capture alignment, as was suggested by the results of the comparison of sequence capture technologies described in Chapter 3, where the amplification based method of RainDance gave significantly higher alignment (80.1% mean) compared to the two hybridisation methods, NimbleGen sequence capture array (51.6% mean) and SureSelect (19% mean).
2. By using Long Range PCR (LR-PCR) only the short sequence which the primers target must match exactly to ensure amplification occurs. Successful amplification can then be confirmed prior to the sequencing by performing gel electrophoresis on the products of the reaction and verifying that a band of approximately the correct molecular weight has been produced. If mutations are present anywhere else in the sequence of the amplicons it will not make a difference to the capture of the sequence, and these mutations should therefore be sequenced by NGS.

A series of primers were designed for LR-PCR that would generate amplicons up to 10kbp in size, which would cumulatively cover the entire region of the *ABCA4* gene, including the 5' and 3' UTRs.

In addition, four primers were designed to amplify two regions in the intergenic sequence surrounding the gene, where the ENCODE data showed high H3K27AC marks. These histone

marks reliably predict transcription factor binding sites that are often associated with regulatory function (Creyghton *et al.* 2010). The promoter regions and the UTRs had not been sequenced in the initial comparison of technologies described in Chapter 3.

In total 24 LR-PCR amplifications were performed for each of 20 DNA samples. Of these 20 samples, 14 were from the cohort of patients with a missing pathogenic variant in the *ABCA4* gene, or from an affected relative of one of these patients. Six of the samples were healthy controls. Amplicons were separated by size and imaged using agarose gel electrophoresis as shown in **Figure 6.12**. Initially amplicons were gel extracted to purify them, however gel extractions did not contain enough DNA to sequence.

LR-PCR Products were Sanger sequenced directly, without agarose gel purification, to confirm that the amplified sequence mapped to the target region of the *ABCA4* gene. Analysis using the UCSC Genome Browser confirmed the amplifications covered the predicted regions (**Figure 6.13**).

All the amplicons from each patient were purified and pooled to an approximately equimolar concentration using SequalPrep plates (as described in section 6.2.3.1 p.222). Pooled amplicons were delivered to the Wellcome Trust Centre for Human Genetics high throughput sequencing group where they were fragmented by sonication, prepared into DNA libraries, and then sequenced on a single flow cell on the Illumina HiSeq 2500.

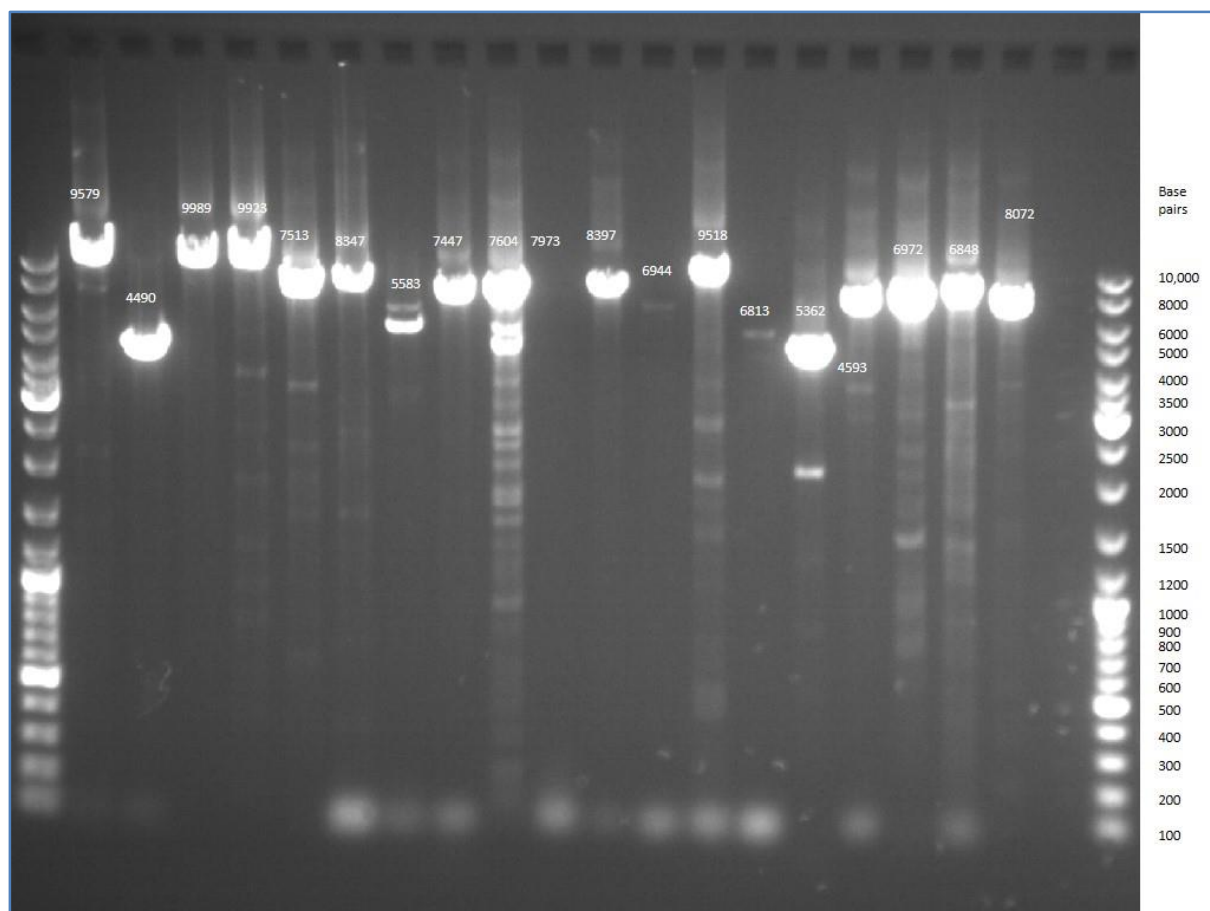


Figure 6.12 Example gel electrophoresis image showing 18 Long Range PCR amplicons covering the *ABCA4* gene. The outside lanes contain DNA marker ladders, and the second well from right is a negative water control. In the 10th sample well from left the reaction has failed. In all other wells an amplicon has been successfully generated at approximately the expected position.

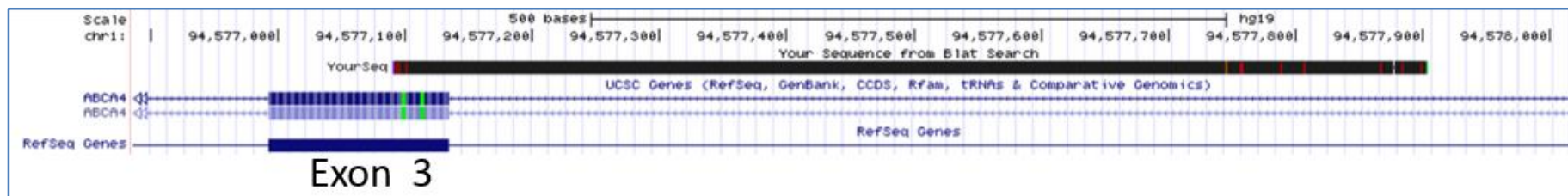


Figure 6.13 Example screenshot from UCSC genome browser confirming an amplification maps to the correct place on the *ABCA4* gene. The black bar shows the Sanger sequence generated from sequencing the amplicon from one end. The amplification was expected to cover the region between exons 1 and 3, and this trace confirms one end of the amplicon extends out from exon 3.

6.3.3.1 Preliminary data from sequencing of LR-PCR products

Preliminary data from the sequencing shows that an average of 90.1% of reads in each sequenced patient sample mapped back to the target region containing the *ABCA4* gene (**Table 6.6**). This target region is approximately 150,000bp in size, which accounts for just 0.0047% of the genome. The high percentage of reads mapping back to the target confirms the experiment was very successful in amplifying and sequencing the *ABCA4* gene.

In future work the data generated will need to be analysed in detail to determine whether variants have been detected through this method that were missed through other methods. In particular, careful analysis of the data will be required to detect evidence of any large scale rearrangements in the *ABCA4* gene can be detected.

The sequencing used paired-end reads, so evidence of a rearrangement might be detected by finding read pairs in which one read maps to a different area of the reference sequence and is in reverse, as shown in **Figure 6.14**. Such reads would likely be rejected by the quality control step of a bioinformatic assembly program which maps reads to a reference sequence.

These programs often create a file containing the discarded reads that were rejected due to their failure to align to a region of the genome, either in full or in part. The reads in these files will therefore need to be analysed carefully to see if some reads map accurately, with a second paired read that does not, which could be evidence of a large rearrangement that could be pathogenic, and account for the failure to find a second pathogenic variant in the *ABCA4* gene in these patients.

Table 6.6 Table of preliminary sequencing metrics data for 14 patients and 6 controls sequenced from LR-PCR amplicons.

Patient ID	Total reads	Reads in target region	
189K0	22,152,013	20,409,509	92.1%
520C7	9,641,873	8,779,200	91.1%
531B2	23,483,323	21,361,761	91.0%
658B6	16,559,211	14,437,860	87.2%
724B6	18,146,688	17,115,837	94.3%
746G4	18,888,413	17,140,731	90.8%
749E4	19,306,827	17,587,017	91.1%
781A3	20,651,714	18,467,062	89.4%
788G4	21,478,243	18,604,406	86.6%
968E3	17,920,146	16,907,195	94.4%
1049H2	20,441,094	17,126,185	83.8%
1062F7	13,214,159	12,038,492	91.1%
1089E2	22,365,160	20,141,892	90.1%
1127F8	18,168,056	15,926,505	87.7%
MSP (control)	22,245,549	20,162,736	90.6%
317A1 (control)	5,625,591	5,282,871	93.9%
832H7 (control)	18,721,196	16,916,631	90.4%
895C2 (control)	22,509,105	20,617,233	91.6%
1089B2 (control)	20,671,457	18,031,113	87.2%
112C4 (control)	21,214,509	19,600,207	92.4%

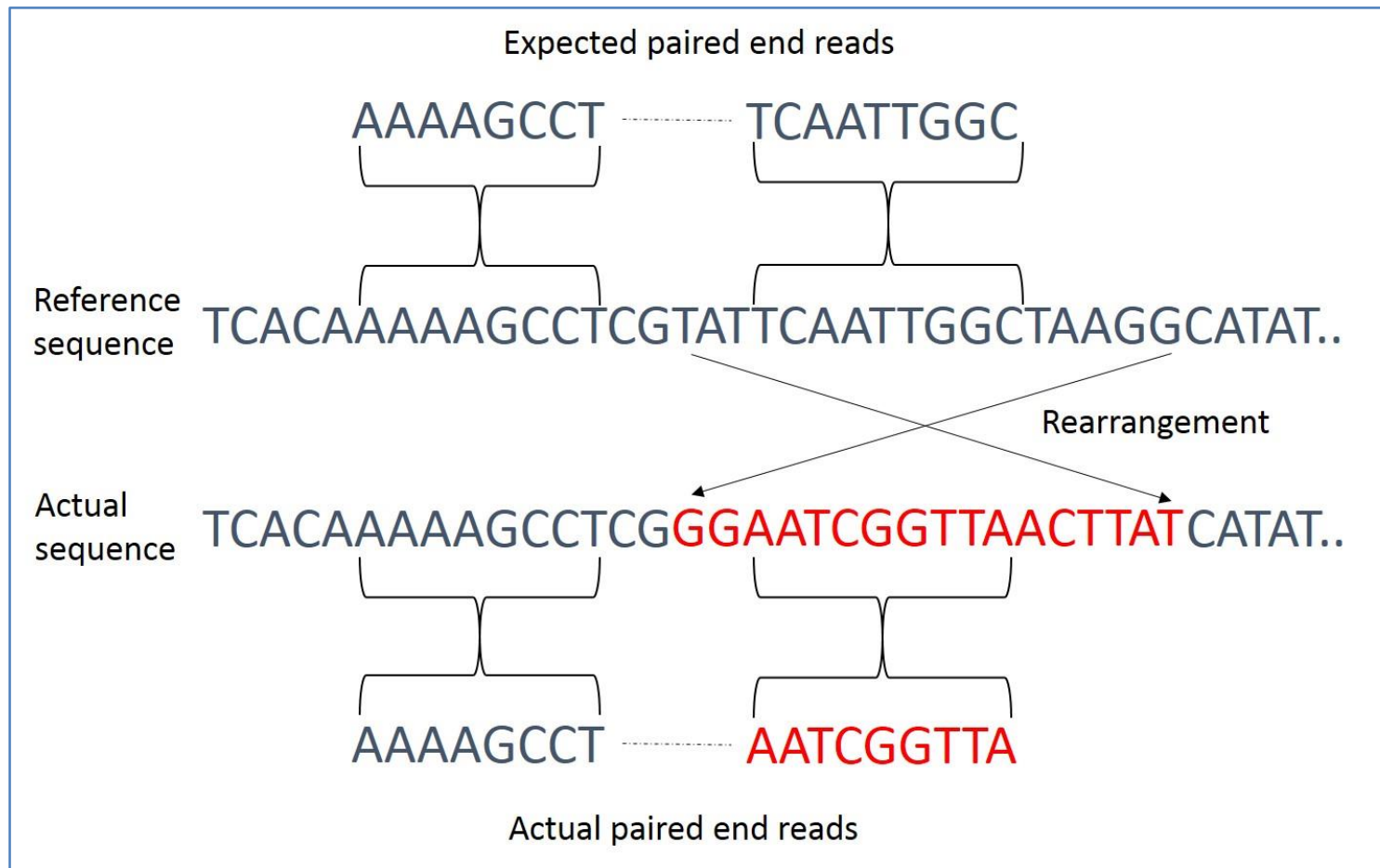


Figure 6.14 The potential effect of a rearrangement on paired end reads produced by NGS. Bioinformatic programs map reads to a reference sequence to compile a continuous consensus sequence. Such programs would therefore expect a pair of reads with the sequence at the top of the image. A rearrangement to the genome which flips part of the sequence (highlighted in red) could end up causing the second read in a read pair to have a very different sequence, which might lead to such reads being erroneously rejected by the quality control function of an assembly program.

6.3.5 Whole genome sequencing of ten patients with Stargardt disease and a single potentially pathogenic variant identified in the *ABCA4* gene

Whole genome sequencing using unchained base reads on self-assembling DNA nanoballs was performed on DNA samples taken from ten patients in whom only a single potentially pathogenic variant in *ABCA4* had been identified. These patients and the phenotypes they were diagnosed with were:

• 189K0 (STGD)	• 749E4 (FF-like)	• 1049H2
• 381E1 (CRD)	• 781A3 (STGD)	(atypical MD)
• 520C7 (STGD)	• 788G4 (STGD)	• 1127F8 (STGD)
• 746G4 (STGD)	• 803H2 (CRD)	

An initial analysis of the data produced was performed using Ingenuity Variant Analysis software with the keywords “Stargardt macular dystrophy”, “macular dystrophy”, and related phenotypes. The software was only able to identify five of the known *ABCA4* pathogenic variants in these patients, as well as four other variants of unknown clinical significance. No pathogenic variants associated with retinal disease were identified in any other genes by the software.

This does not fully preclude the possibility that a variant in another gene is causing the condition of the patients, however it strongly suggests that the as-yet-unidentified pathogenic variant or variants responsible for the condition of the patients have not been previously reported, whether they are in *ABCA4* or another gene.

Further analysis of the data was performed manually by searching for variants that were located close to or inside the novel exons reported by Braun and colleagues, or by Farkas and colleagues in the retinal transcriptome (Braun *et al.* 2013, Farkas *et al.* 2013). Two different potentially pathogenic variants were identified, one present in 2 patients; 189K0 and 1127F8, and another present in 2 patients; 803H2 and 381E1.

6.3.5.1 Example of a cryptic exon

The previous genetic testing performed in patient 189K0 had identified a c.3364G>A (p.Glu1122Lys) substitution in patient 189K0 which was predicted to be Probably damaging by PolyPhen2 (score 1.000), Deleterious by SIFT (score 0.0) and Disease causing by MutationTaster (p-value 1), with a 0.00002471 allele frequency in ExAC. In patient 1127F8 a c.5461-10 T>C substitution of uncertain clinical significance had been identified (Maugeri *et al.* 1999, Rivera *et al.* 2000). The variant causes no significant changes in splice prediction programs, but appears at 0.0002231 allele frequency in the ExAC database.

These two patients were found to have one of the same intronic pathogenic variants reported by Braun *et al.* (2013) in the exon annotated in that paper as 36.1 (see **Figure 6.15 A**). This variant is a c.5196+1137G>A substitution on chromosome 1, located three base pairs 3' to the splice site of the cryptic exon that exists in an alternate transcript of *ABCA4* which the authors report to be pathogenic by increasing the strength of this splice signal, thereby increasing the level of transcription of the cryptic exon 36.1. The variant is marked by the red arrow in **Figure 6.15 A**. The presence of this second potentially pathogenic variant in the *ABCA4* gene of each patient provides a possible molecular diagnosis for each of their conditions.

6.3.5.2 Example of possible alternative splicing

Patients 803H2 and 381E1 were siblings in whom a deleterious nonsense mutation c.655A>T (p.Arg219X) had previously been identified. This nonsense mutation did not appear in the ExAC database. Analysis of the WGS data identified a c.859-637T>C substitution in intron 7 (**Figure 6.15 B i**)). This variant has does not appear in the ExAC database, and appears to be located in an intronic region of *ABCA4* according to the most recent RefSeq human genome build, however new data annotating the retinal transcriptome, published during the course of these experiments (Farkas *et al.* 2013), suggests that the variant may be located in an alternatively spliced exon.

Figure 6.15 B) i) shows the retinal transcriptome data published by Farkas and colleagues as a track on UCSC genome browser. The transcriptome data in **Figure 6.15 B) i)** contains a green bar in the centre of the image which indicates the presence of an exon junction identified by the RNA-seq Unified Mapper (RUM) program (Grant *et al.* 2011).

It is unknown whether this indicates the presence of a cryptic exon in this location, as the retinal transcriptome does not annotate this exon junction as being part of an inferred internal exon.

Figure 6.15 B) ii) shows the nearest inferred internal exon to the location of the variant. It may be that there is an alternative transcript of the gene in which this exon extends out to the position of the exon junction indicated by the RUM high quality junction where the variant is located.

If this is the case it would make the c.859-637T>C variant an exonic variant that could have an effect on coding, rather than a benign intronic variant as previously thought. Further work must be performed to discover whether this variant does in fact cause a change in the transcription of that exon, which will require testing of gene expression in *ABCA4* cDNA.

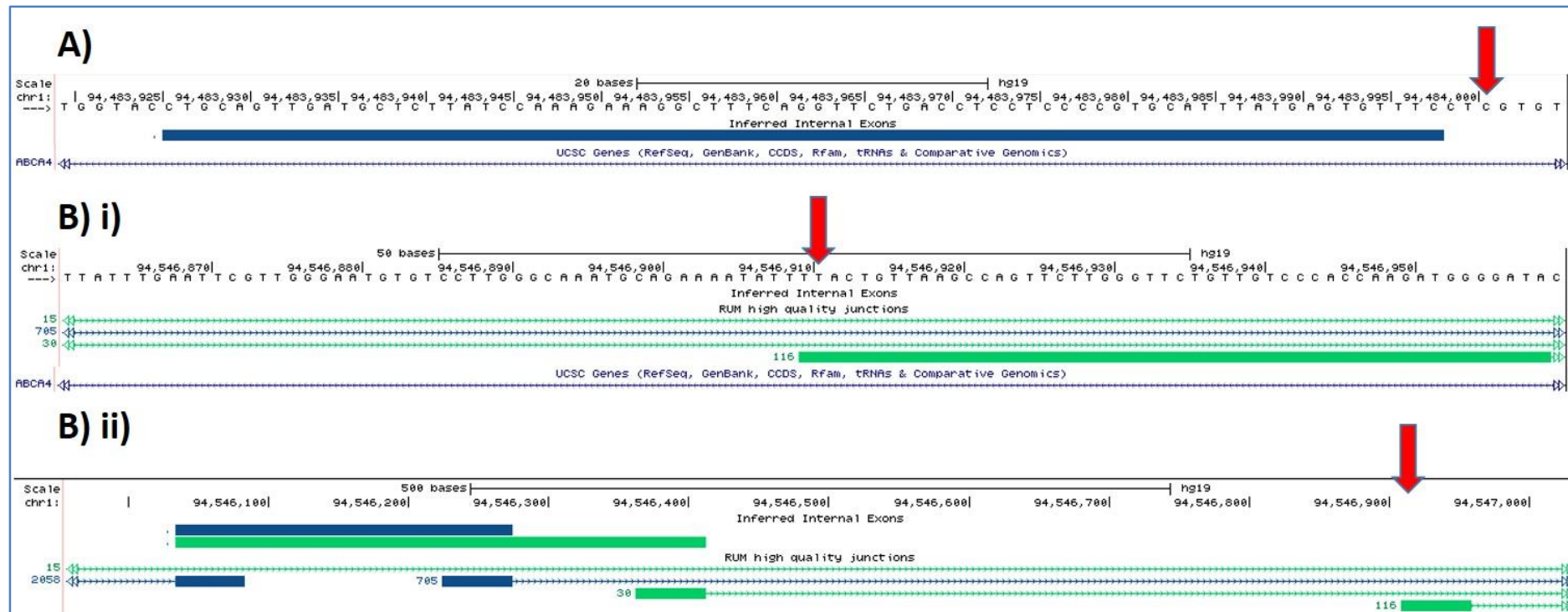


Figure 6.15 Screenshots of the UCSC genome browser displaying data from the retinal transcriptome paper by Farkas *et al.* (2013) as an overlaid track, and the previously annotated positions of UCSC genes on the bottom track. The red arrows indicate the positions of variants uncovered through whole genome sequencing **A)** A c.5196+1137G>A substitution in patients 189K0 and 1127F8. **B) i)** A c.859-637T>C substitution in patients 803H2 381E1. **B) ii)** The same variant when viewed whilst zoomed out to show the nearest inferred internal exons.

6.3.6 *ABCA4* cDNA analysis in patients using mRNA from lymphoblastoid cell lines

In order to investigate the expression of *ABCA4* in single variant patients, as described in section 6.1.5 p. 214, an attempt was made to extract mRNA from lymphoblastoid cell lines and reverse transcribe this to cDNA. Cell lines were made from whole blood samples of patients 520C7, 531B2, 658B6, and 749E4 as described in section 6.2.4 p.224. mRNA was extracted using TRIzol and an acid guanidium thiocyanate-phenol-chloroform extraction. cDNA was reverse transcribed and successive PCR and nested PCR amplifications were performed over two areas of the *ABCA4* gene. The products of these nested PCRs were run on an agarose gel and showed bands present in the expected locations (**Figure 6.16**).

The bands produced by the nested PCRs in these experiments included amplifications that covered exons 36-37. The nested PCR products covering these exons were Sanger sequenced, however in each case the sequence generated matched the reference sequence. The cryptic exon 36.1 in intron 36 described by Braun and colleagues (2013) was not present in the products.

Repeated attempts were made to perform nested PCRs across other exons of *ABCA4*, and also to amplify over exons 36 and 37 in patients 189K0 and 1127F8, in whom the whole genome sequencing had identified a pathogenic variant in exon 36.1, identified by Braun and colleagues.

However, these attempts all failed to produce amplifications, possibly owing to the low level of *ABCA4* expression in blood and the difficulty of successfully performing nested PCR. These experiments need to be repeated in future work to discover whether the pathogenic exon 36.1 exon can be detected in cDNA transcripts in these patients.

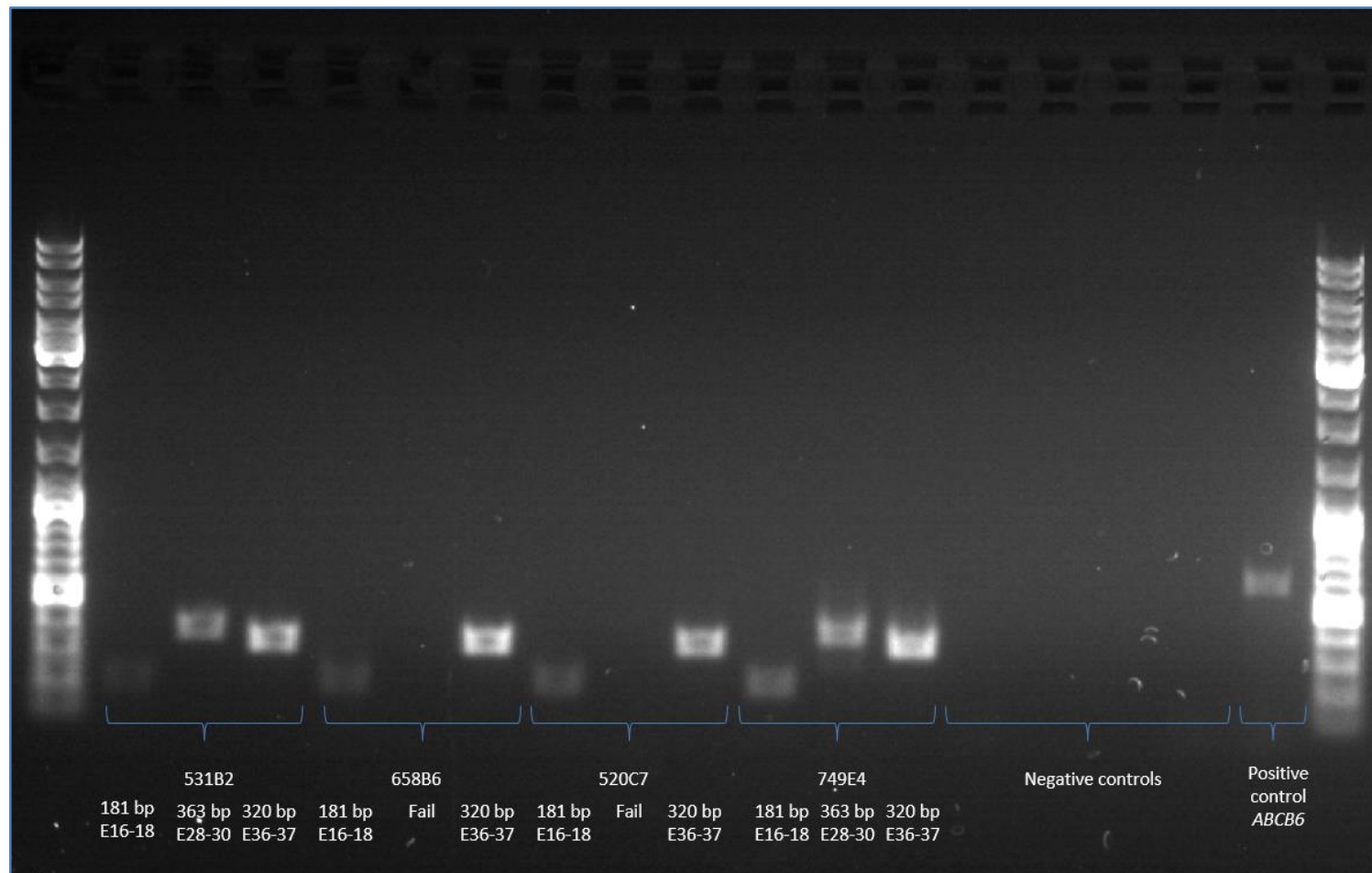


Figure 6.16 Successful nested PCR amplifications of *ABCA4* cDNA. Agarose gel electrophoresis image showing successful nested PCR amplifications across three regions of the *ABCA4* gene in four different patients. The amplification covering exons 28-30 has failed in two of the patients.

6.4 Discussion

The aims of the work described in this chapter were to identify a molecular diagnosis in two cohorts of patients with macular dystrophies. In one cohort the patients had a clinical diagnosis of Stargardt disease. The other cohort consisted of patients with non-Stargardt MD phenotypes. Sanger sequencing had identified just a single potentially pathogenic variant in the *ABCA4* gene in each one of the patients of each cohort.

Four different approaches were used to try and identify either a second pathogenic variant in the *ABCA4* gene, or a mutation in another genetic locus which could be responsible for the condition. In each case the experiments provided data that either identified a molecular diagnosis for one or some of the patients, or provided a foundation on which further research can be performed.

1. The targeted sequencing identified a pathogenic mutation in the *CRX* gene of patient 245F6 who had a clinical diagnosis of Bull's eye maculopathy, demonstrating the ability of NGS to provide an unexpected molecular diagnosis in conditions that show phenotypic heterogeneity.
2. The analysis of introns using the data from Farkas and colleagues (2013), and from Braun and colleagues (2013), identified intronic variants in *ABCA4* that cause pathogenicity by activating a cryptic exon, or by possibly changing the splicing of an existing exon.
3. The sequencing of LR-PCR amplicons has created a resource of high quality data that can be mined for pathogenic variants in future analyses.
4. The RT-PCR experiments were able to amplify cDNA from the *ABCA4* gene, but the results were not consistent enough to pursue. Future work might include using different tissue sources rather than cell lines or peripheral blood which may amplify more efficiently as expression may be higher.

In total a potential molecular diagnosis was found for two of the fourteen patients with Stargardt disease, and three of the eight with other macular dystrophy phenotypes, shown in **Table 6.7**.

The analysis of NGS data to find pathogenic variants can be complex, as was previously discussed in section 4.4 p.160. In the investigations described in this chapter the data produced from the targeted resequencing of patients with maculopathies was further analysed. In particular, the analysis was extended to examine 73 other retinal genes associated with retinal dystrophies, which could potentially be the cause of the condition in the patients.

In patient 245F6 the analysis of this NGS data uncovered an R41W mutation in the *CRX* gene, which has been previously reported to cause cone-rod dystrophy. The mutation was then found in the affected sister, as well as the father who had not reported problems with vision, but on autofluorescence imaging was shown to have patchy degradation of the macula.

The discovery that this mutation was responsible for the condition of the patient was unexpected, owing to the phenotype observed, and is a good example of the power that NGS has to produce medically useful information. In this instance the change of the diagnosis from a suspected recessive *ABCA4* maculopathy to a dominant *CRX* mutation, caused a change in the genetic counselling to the patient, as there was now a 50% recurrence risk, compared to the relatively low risk from recessive mutations in *ABCA4*.

The finding that the 36.1 exon variant was only identified in the data produced by the RainDance and 454 technology combination in patient 189K0 was an interesting discovery, as it suggested that there might be a bias in hybridisation capture methods which prevented the intronic variant from being captured and sequenced by the NimbleGen sequence capture array.

Most intronic DNA is non-functional which means that many mutations occurring in these regions can accumulate without having any impact on the survival of the organism, as opposed to exons where mutations are more likely to have a deleterious functional effect. This causes the nucleotides in exonic regions to be more highly conserved by evolution, and as a result intronic regions show greater variation than exonic regions of the genome.

Table 6.7 Patients in whom a potential molecular diagnosis has been identified from the investigations in this chapter. The newly identified pathogenic variants are highlighted in yellow.

Patient ID	Phenotype	Potential pathogenic variants identified		
		Gene	Nucleotide change	Protein change
245F6	BEM	<i>ABCA4</i>	c.4685T>C	p.Ile1562Thr
		<i>CRX</i>	c.121C>T	p.Arg41Trp
381E1	CRD (historical STGD)	<i>ABCA4</i>	c.655A>T	p.Arg219X
		<i>ABCA4</i>	c.859-637T>C	Alternative splicing?
803H2	CRD (historical STGD)	<i>ABCA4</i>	c.655A>T	p.Arg219X
		<i>ABCA4</i>	c.859-637T>C	Alternative splicing?
189K0	STGD	<i>ABCA4</i>	c.3364G>A	p.Glu1122Lys
		<i>ABCA4</i>	c.5196+1137G>A	Exon 36.1?
1127F8	STGD	<i>ABCA4</i>	c.5461-10T>C	Splicing
		<i>ABCA4</i>	c.5196+1137G>A	Exon 36.1?

As a result, capture designs based on the reference sequence will produce probes that match less exactly to the target in intronic regions, and so will capture these sequences less efficiently than more highly conserved exonic regions.

In addition, during the assembly of continuous sequences from individual reads, a reference sequence is often used as a guide. This can also lead to errors, as reads containing intronic sequence which show substantial variation from the reference sequence may be falsely identified as erroneous and discarded by the assembly algorithm. Any of these reasons may explain why the NimbleGen sequence capture array method was less efficient than the Raindance method.

This was part of the reason why it was decided that long range PCR would be performed as a capture method for targeted sequencing. The initial experimental design for the LR-PCR and sequencing also included gel extraction and purification of bands containing amplified sequence. This would theoretically have meant that only the target sequence would have been sequenced, so all reads generated by NGS could be assembled using a reference sequence. If any reads were rejected by the assembly software and were not obviously sequencing errors, they could be analysed to determine whether they contained a large variation to the sequence such as a large re-arrangement or large deletion.

The attempts at gel extraction and purification failed to yield enough DNA for sequencing however, so instead the products were directly purified and pooled before sequencing using SequelPrep plates. This meant that the sequenced data contained DNA from other regions of the genome which were present in the original template DNA used for amplifications.

Despite the lack of gel extraction, the preliminary results produced from the sequencing of the amplicons from 20 patients using this method showed that the LR-PCR was a very efficient sequence capture and enrichment method with an average of 90.1% of sequence reads from each patient sample mapping back to the target region containing the *ABCA4* gene. The average 9.9% of reads that did not map to the target sequence, would have been from template

genomic DNA used in the LR-PCR, which contains the remaining 99.9953% of the genome (the vast majority of it).

Further analysis will need to be performed on the data produced from this sequence enrichment method. The variant called files (VCFs) generated from the assembled sequence must be analysed to determine whether there are intronic variants present that might create cryptic exons, either in the exon 36.1 described by Braun and colleagues (Braun *et al.* 2013), or in other introns of the gene. Careful analysis of the data must also be performed to determine whether any of the rejected reads actually show evidence of a large rearrangement.

Although it was not possible to analyse the data generated from LR-PCR in the time available, data from whole genome sequencing was produced and analysed over the course of these investigations. The data produced from WGS by Complete Genomics was delivered in their own data file format termed “master variant files” (masterVar) which represent an improvement on the variant called files (VCFs) that have become the standard in NGS. There are more annotations and the columns are labelled with easier-to-understand, more user friendly headings. More significantly, instead of numbering variants by the start and end position of the variant, these data files label the gaps between the variants. This allows indels to be more accurately recorded, and for the information to be processed more easily by bioinformatic pipelines.

The data produced by the WGS of ten patients identified several potentially interesting variants. In two of the patients, 189K0 and 1127F8, the pathogenic variant in exon 36.1 was identified confirming the finding of this variant in the targeted capture data from 189K0 and for the first time in 1127F8.

Using data from the retinal transcriptome published by Farkas and colleagues it was possible to identify other variants that appeared in other exons that were not annotated on the most recent Reference Sequence build. This included a c.859-637T>C substitution which appeared to create a new exon junction that greatly increases the size of an existing exon. This shows the

importance of matching genomic information with high quality transcriptomic information derived through cDNA sequencing. Further testing will be required to confirm whether this change to the size of the exon is actually being expressed in mRNA.

For this reason, attempts were made to amplify cDNA from the *ABCA4* gene in these patients. mRNA was extracted from lymphoblastoid cell lines derived from patient blood samples. This was reverse transcribed into cDNA, and amplified by PCR followed by nested PCR. Sanger sequencing was performed on the products and in each case no changes from the reference sequence were found. Successive attempts were made to amplify other areas of the gene, or to repeat the experiments for samples from other patients, but these attempts failed, possibly owing to the low level of *ABCA4* expression in lymphoblastoid cell lines, as evidenced by the need to perform nested PCR to amplify sequences from the gene.

By attempting to analyse *ABCA4* cDNA it was hoped that it would be possible to sequence the regions of the gene where the already identified pathogenic mutations are located. If a patient has compound pathogenic variants on each allele of the *ABCA4* gene, but neither of these cause nonsense mediated decay then it should be possible to see each of these mutations as a heterozygous variant on a Sanger sequencing trace.

Such a result would strongly suggest that the search for a missing pathogenic variant in *ABCA4* should continue, as opposed to searching in other genes. Further work must be performed on the mRNA from these patients to produce such a result. In particular mRNA must be extracted from other sources such as skin or saliva where the expression level of the gene may be higher, creating a higher concentration of product from reverse transcription.

6.4.1 Further work and conclusion

The work described in this chapter was able to successfully identify the missing pathogenic variants in two of the fourteen patients with Stargardt disease who lacked a secure molecular diagnosis. In contrast to Braun *et al.* (2013) who reported 50% of the cases with mutations

affecting alternative transcripts, in our data set 12/39 cases are yet to have a second pathogenic mutation found.

In the introduction to this chapter a list of possible explanations for why a second pathogenic variant had not yet been identified in these patients was proposed. The possibilities that the condition of the patients was being caused by intronic variants, or variants in other genes were both investigated in this chapter with some limited success. It seems probable that the remaining missing variants would be caused by the other explanations for which there was not time to investigate; mosaicism, miRNA binding, trinucleotide repeat expansion, other repeats, exonic splice enhancers or silencers, or epigenetic effects. A brief analysis of the whole genome sequencing data to search for pathogenic variants in other genes yielded no results, however this does not preclude the possibilities that these patients have a pathogenic variant in another gene associated with retinal dystrophy that has not previously been reported, or indeed a mutation in an as-yet-unidentified gene.

Future work on the genetic causes of Stargardt disease, and indeed other recessive conditions where pathogenic variants are missing, will have to focus on investigating these other possible explanations. In some cases the technology required to adequately investigate these possibilities requires further development, such as in the case of repeat expansions which may not be accurately detected by current NGS methods that rely on short reads. In other cases, such as exonic splice enhancers and silencers, our understanding of these phenomena must improve, and the possible DNA sequences that perform such purposes must be catalogued in a database to better allow their detection using bioinformatic programs.

The work in this chapter was also able to provide a secure molecular diagnosis for one of the eight patients with a macular dystrophy phenotype other than STGD, patient 245F6 in whom a dominant mutation in the *CRX* gene was identified, and a potential molecular diagnosis for two others, patients 381E1 and 803H2, in whom a second potentially pathogenic intronic variant in the *ABCA4* gene was identified.

This leaves at least five patients in whom a molecular explanation for their condition is still required. Given the genetic heterogeneity of macular dystrophies there is a strong possibility that the condition of these patients is being caused by mutations in another genetic locus, and as the number of genes known to have an association with retinal diseases has so far continued to increase, it may well be possible that the pathogenic mutations reside in a gene that is not currently known to cause macular dystrophy.

Future work must to continue to identify genes that are involved in retinal function, and identify different mechanisms by which a mutation in a gene can cause a pathogenic effect.

CHAPTER 7

General discussion

The overall aim of the work described in this thesis was to apply next-generation sequencing technologies for use in the molecular diagnosis of genetic diseases. This project set out to compare and validate platforms using known patient variants and mutations, and then moved on to testing patients with an unknown diagnosis. Initially, targeted capture and NGS were used, and then exome and whole genome sequencing.

At the time this work was started in October 2010 very few papers had been published showing the application of NGS for use in a clinical diagnostic setting for genetic diseases. Since there was so little data on accuracy and reliability of the technologies (both critical for clinical diagnostics), the initial investigations were performed to determine whether these technologies could realistically be introduced into clinical practice and how this might change the concepts of genetic testing in patients. Each of the technology combinations was found to perform well, identifying the majority of variants detected by Sanger sequencing, and the results were similar to those reported by other research groups.

One of the earliest papers to compare NGS platforms was published just a few months before these investigations began. The authors used just one sequence capture technology: the NimbleGen sequence capture array, and compared sequencing of the captured samples on the 454 GS FLX and the Illumina GA (II) platforms (Hoppman-Chaney *et al.* 2010). Their results showed a higher average depth of coverage for the Illumina platform compared to the 454, as was found in this work, however the depth of coverage for NimbleGen sequence capture array and 454 sequencing was higher in this work: >88% bases covered at $\geq 20\times$ depth, compared to just >30% bases covered at $\geq 20\times$ depth reported by Hoppman-Chaney *et al.* (2010).

Having evaluated the ability of NGS and sequence capture technologies to provide medically useful information, the experiments in this project then moved forward to test patients without a molecular diagnosis. The initial decision of which NGS platform to use for further experiments was ultimately decided by the bioinformatic support provided by 454. At that time the Illumina platform provided very little bioinformatic support, and required specialist bioinformaticians and

computer software programs. In the early days of developing NGS for clinical diagnostics this was not easily available and the Roche integral software was an attractive option.

Perhaps for this reason, the Roche 454 was also used by other groups who published papers at the time such as Neveling *et al.* who later captured 111 genes known to be associated with retinitis pigmentosa on the NimbleGen sequence capture array (Neveling *et al.* 2012).

Since then Illumina have provided a cloud based bioinformatic pipeline that performs many of the functions the GS Mapper software provided for 454. Illumina have also vastly improved the throughput of their sequencing machines, and importantly a tranche of bioinformaticians around the world are being trained to analyse the data. As a result, the Illumina platform was chosen for the later experiments in this project. This is clearly also the experience of other scientists as the Illumina platform is currently the “market leader” and the platform of choice for many international studies.

One of the first papers to use the Illumina platform to sequence retinal disorder captured 51 genes in five patients using the NimbleGen sequence capture array and sequenced on the Illumina GA (II) platform, finding known pathogenic mutations in two of the patients and a novel potentially pathogenic variant in a third (Simpson *et al.* 2010).

A recent comparison of the Roche 454 GS FLX and Illumina HiSeq 2000 by Li *et al.* reported a similar observation to this work: Illumina gave increased depths of coverage, which gave greater sensitivity in the detection of minor variants, and fewer false positives (Li *et al.* 2014).

Another paper compared the Roche 454 GS FLX, the Illumina HiSeq 2000, and another commonly used NGS system not used in the work described in this thesis, the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system by Applied Biosystems. Their analyses discovered a GC bias in the Illumina and SOLiD platforms, and they also suggested that sequencing human genomes with more than a single platform and multiple libraries could be beneficial when a high level of accuracy is required (Ratan *et al.* 2013).

The SOLiD system was used for a screen of IRD conditions that captured 105 genes using SureSelect, detecting pathogenic mutations in 55% of RP cases and 80% of Usher or Bardet-Biedl syndrome cases (Glockle *et al.* 2014).

The largest next-generation sequencing facility in the world is the Beijing Genomics Institute (BGI) who reported in 2012 that their sequencing facility is made up of 137 Illumina HiSeq 2000 machines, 27 SOLiD platforms, 1 Ion Torrent PGM, 1 Illumina MiSeq, and 1 Roche 454 GS FLX (Liu *et al.* 2012). Their conclusions on comparing the different platforms available were that HiSeq has the biggest output and lowest reagent cost, SOLiD has the highest accuracy, and the Roche 454 has the longest reads (Liu *et al.* 2012). Given their choice of sequencing machines it would appear that cost remains the biggest determinant of which platform to use for the majority of sequencing projects.

The majority of the papers published however have reported work performed in a research setting, and the work undertaken in this project was some of the first to implement NGS into a clinical diagnostic setting. As a result of research projects including the work presented here, the NHS Oxford medical genetics laboratories now offer NGS screening for:

- 8 candidate genes for arrhythmogenic right ventricular cardiomyopathy (ARVC).
- 99 candidate genes for ataxia.
- 28 candidate genes for familial dilated cardiomyopathy (DCM).
- 16 candidate genes for familial hypertrophic cardiomyopathy (HCM).
- 28 candidate genes for Joubert syndrome.
- 19 nuclear candidate genes for mitochondrial disorders.
- 115 candidate genes for inherited retinal disorders.

Information from: <http://www.ouh.nhs.uk/services/referrals/genetics/genetics-laboratories>

In other centres around the UK NGS screens have been introduced for conditions such as hereditary spastic paraplegia (HSP), neuropathies, dystonia, and dementia. The majority of the

clinical screens that have been introduced still rely on targeted capture of specific genes. For example, at a similar time as the work in Chapter 4 was published, another group in the UK reported the use of Agilent SureSelect and SOLiD sequencing for a panel of 105 retinal genes known to be mutated in patients with Leber congenital amaurosis, macular dystrophy, achromatopsia, and Usher and Bardet-Biedl syndromes (BBS). Their screening method gave a detection rate of 50-55%, which represented a significant increase of the detection rate compared to Sanger sequencing (O'Sullivan *et al.* 2012). The authors also note that the NGS panel they use can be applied to a broader range of patients, whereas Sanger sequencing was only offered to around a third of patients with either autosomal dominant or X-linked forms of disease (O'Sullivan *et al.* 2012).

7.1. The progression of DNA sequencing technologies

In the years since the work described in this thesis began the throughput of NGS technologies has continued to increase, and the cost of DNA sequencing has fallen further still, as shown in **Figure 7.1**. Between 2001 and 2007 the decrease approximately followed Moore's law, a trend first described by Gordon Moore of Intel corporation, describing how in the computer industry computing power would double every two years for the same amount of cost. The yellow arrow in **Figure 7.1** marks the acquisition of 454 corporation by Roche in 2007 and the start of the NGS era. At this point the cost of sequencing fell dramatically for a period of four years as competing NGS technologies drove the price down.

The red arrow in **Figure 7.1** marks the point in time at which these investigations began, and at this point the cost of sequencing a single human genome was still approximately \$100,000, which made it prohibitively expensive for routine clinical diagnostic use, and necessitated the use of sequence capture technologies to make the investigations described in this thesis economically feasible. As the cost of sequencing fell further still, it became economical to sequence larger targets, and eventually it was more practical to perform exome sequencing, rather than constantly create new target designs.

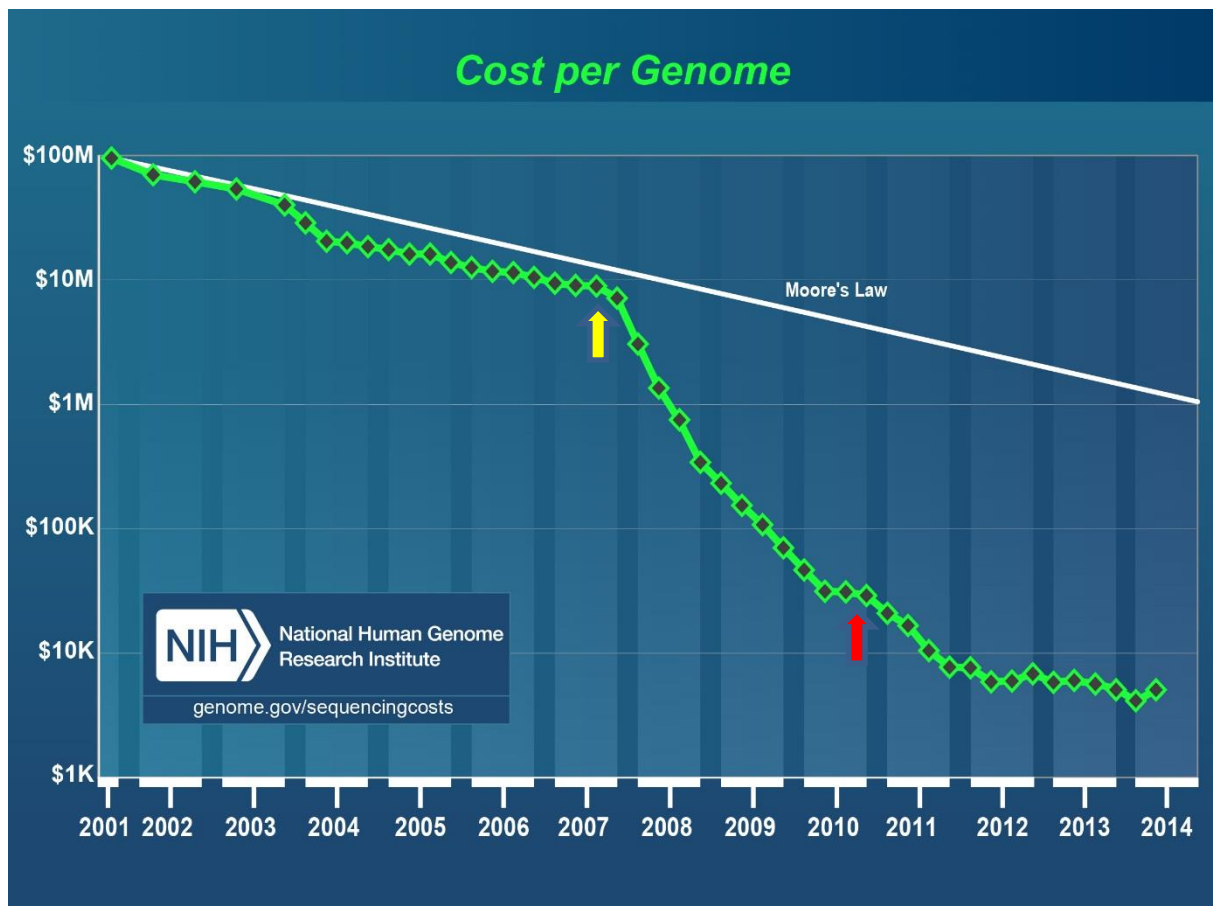


Figure 7.1 The cost of genome sequencing over the previous 14 years. The yellow arrow indicates the point at which next-generation sequencing technologies first became available. The red arrow indicates the point at which the work described in this thesis began. Figure adapted from: www.genome.gov/sequencingcosts

Sequencing the exome should in theory sequence every gene in the genome, so that for many conditions such as inherited retinal dystrophies (IRD), as genes are continually being identified as having an association with the disease (shown in **Figure 1.27** section 1.8.1 p. 65) they can be investigated. If a molecular diagnosis is not found, the data will be available in future to look at other genes, as associations between those genes and the disease is discovered, without the need to redesign capture probes.

A 2012 investigation of patients with paediatric-onset neurodevelopmental disease showed that exome sequencing improved diagnosis compared to previous genetic testing using Sanger sequencing of the known genes, and was able to not only identify variants in other genes, but to also identify 22 new gene associations with the disease, something that targeted capture by definition cannot do. In addition the sequencing uncovered pathogenic variants in 10 patients that were associated with a different disease to the original diagnosis, changing the diagnosis of the patient and the genetic counselling they received (Dixon-Salazar *et al.* 2012).

Exome sequencing has limitations of its own though. In a comprehensive comparison of exome capture technologies produced by two different manufacturers Asan and colleagues (2011) showed that there were marked differences in what was captured and sequenced by NimbleGen and SureSelect capture technologies. The capture designs themselves had significant differences in what was actually targeted. NimbleGen captured 34.1Mb of sequence, while Agilent SureSelect captured 37.6Mb of sequence. 30Mb of this sequence was shared by both designs, giving 4.1Mb of sequence that was unique to NimbleGen and 7.6Mb that was unique to Agilent SureSelect. The sequence that was unique to each capture design included 305 protein coding genes that were unique to NimbleGen and 1,316 that were unique to Agilent SureSelect. Their results show that there are substantial differences in what is sequenced when purchasing a capture design for the “exome” from different commercial manufacturers (Asan *et al.* 2011).

In addition to the differences in the target design, Asan and colleagues found that there were other sources of bias and error in capture designs. All three capture designs relied on

hybridisation methods to capture DNA, and this resulted in a bias towards capturing reference sequence which bound with greater affinity to the target probes. There was also a significant bias towards regions higher in GC content, due to the additional hydrogen bonds between the pairing of these molecules giving them increased stability, which is an issue that will arise with any kind of hybridisation based targeted capture method, and not exclusively with exome capture (Asan *et al.* 2011).

This inherent bias of hybridisation based capture methods prompted the design of the long range PCR amplifications of the *ABCA4* gene described in chapter 6, and in future work it may be possible to detect variants from the data generated by this sequencing which have been missed by previous targeted sequencing attempts that relied on hybridisation based methods.

Towards the end of the investigations in this project the cost of sequencing had fallen so much that whole genome sequencing had become economical enough to be used as a method for detecting pathogenic mutations. Whole Genome Sequencing has the added benefit of removing all the biases and sources of error produced by using any kind of capture method.

A recent project compared exome capture using NimbleGen sequence capture array, and the two most recent versions of SureSelect kits, to whole genome sequencing performed by Complete Genomics, and on the Illumina X Ten sequencing platform, and found that whole genome sequencing actually provides higher coverage of coding bases than exome sequencing for comparable sequencing depths (Lelieveld *et al.* 2015).

Another recent study showed that WGS was able to identify coding variants that were missed completely by exome sequencing, suggesting that the enrichment of coding regions can actually introduce a bias that prevents these regions from being sequenced (Belkadi *et al.* 2015).

However, there remain biases in the data produced by WGS as well, caused by the different available sequencing methods themselves. For example, the ligation based sequencing provided by Complete Genomics, which was used for some samples in the Stargardt cohort, is

a sequencing method that uses very short reads, which can lead to inaccuracies when trying to detect repeat regions and large re-arrangements.

The move from targeted sequencing, to exome sequencing, through to whole genome sequencing that was undertaken in these investigations was mirrored by the NGS-using research community as a whole over the course of time in which this project took place. Each of these transitions allowed a far greater amount of sequence data to be generated.

This presented a challenge in itself, as although the chance of sequencing a pathogenic variant increased, the amount of benign variants and incidental findings in the final data both increased as well. This led to a greater need for bioinformatic filtering of the data, and placed a greater strain on the computer infrastructure required to store and analyse the vast amounts of sequence data produced.

7.2 Data analysis

At the time the work in this project began very few papers on NGS for clinical diagnostic use had been published, and in the few that had been there was considerable variation in the pipelines used for analysing data.

In many papers variants were presumed not to be pathogenic and filtered out if they were listed in dbSNP (Chou *et al.* 2010, McBean *et al.* 2015). In the work reported in this thesis variants were not filtered out if they were present in dbSNP, as many pathogenic variants had dbSNP entries with no description of their clinical significance.

The use of pathogenicity prediction programs such as PolyPhen2 (Adzhubei *et al.* 2010) and SIFT (Kumar *et al.* 2009) has become commonplace in the analysis of NGS data, and they often form part of the analysis pipelines of NGS projects, along with other pathogenicity prediction programs such as Mutation Assessor (Reva *et al.* 2011), MutationTaster (Schwarz *et al.* 2010), Multivariate Analysis of Protein Polymorphism (MAPP, Stone and Sidow 2005), Protein

Annotation Through Evolutionary Relationship (PANTHER, Mi *et al.* 2013), and Condel (Gonzalez-Perez and Lopez-Bigas 2011).

These programs rely on different methods to predict the effect of a nucleotide substitution on amino acid function, which in most cases include evaluating evolutionary conservation scores of each amino acid. They can sometimes also include the predicted effects on the physicochemical constraints of the amino acid, such as charge and polarity.

Although prediction programs provide a starting point for analysis of data they are not 100% accurate, and so even after they are used for a filtering step more analysis must be performed on the remaining data. A variety of methods for proceeding from this point have been described by different groups. Wei and colleagues reported in 2011 that they used NCBI dbSNP or HapMap databases to determine previously reported mutations, and determined pathogenicity using HGMD or previous reports in the literature (Wei *et al.* 2011). A similar procedure was reported by Blue and colleagues (Blue *et al.* 2014).

A recent paper by Watson and colleagues described their methods for using next-generation sequencing to screen for pathogenic variants in patients with retinal dystrophies (Watson *et al.* 2014). They used a targeted SureSelect capture design for 200 genes which at the time were listed on RetNet as having an association with retinal dystrophies, with sequencing performed on the Illumina GA IIx sequencer (Watson *et al.* 2014).

Their analysis pipeline is shown in **Figure 7.2** and involves annotation of variants using Alamut Batch, screening out synonymous variants, variants more than 5bp from a splice junction, and those with a minor allele frequency of more than 0.01 on dbSNP (Watson *et al.* 2014). Their pipeline shows many similarities to that described in this project, with variants annotated using pathogenicity prediction programs, then filtered out if located outside exons or splice junction,

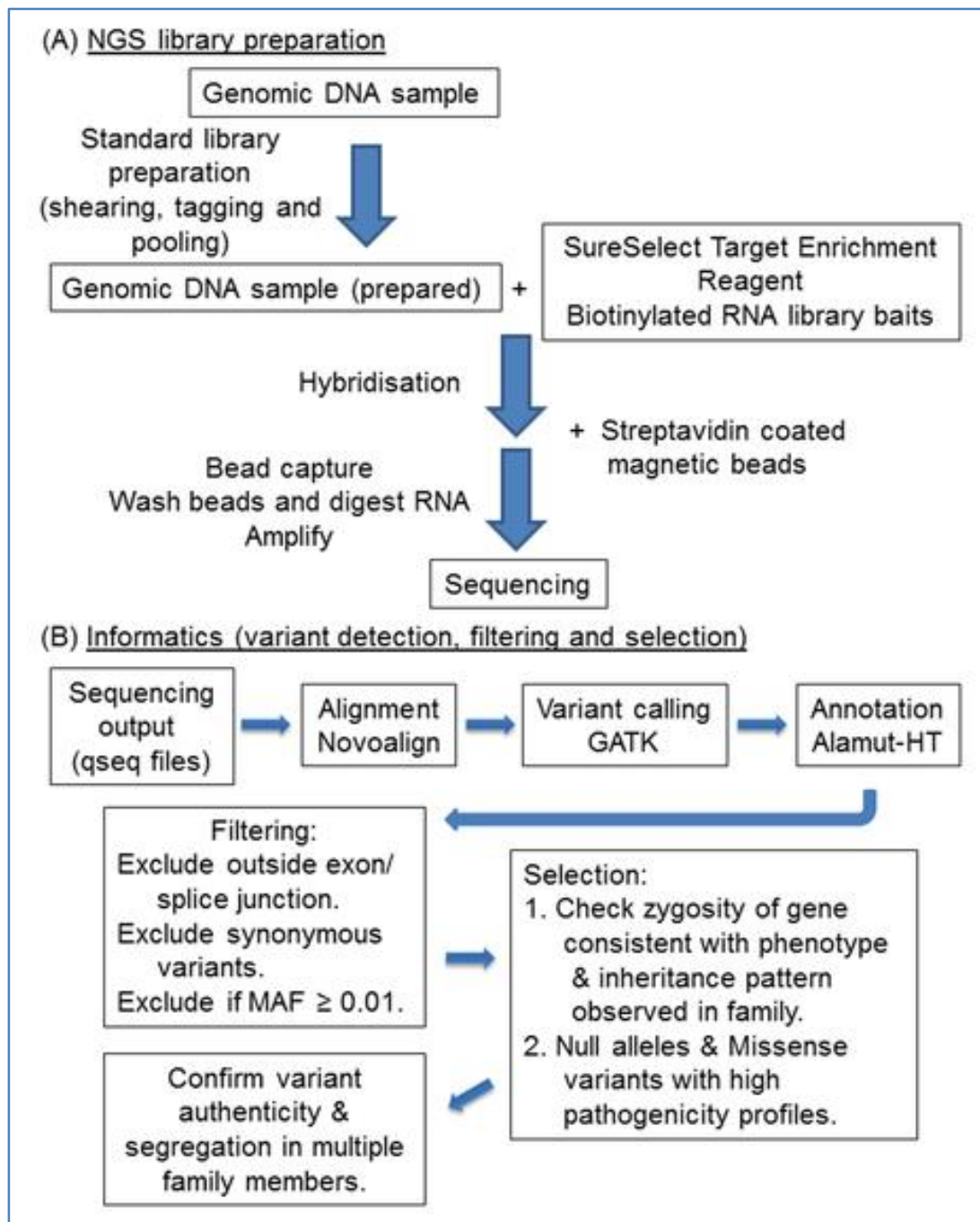


Figure 7.2 Schematic for next-generation sequencing and variant detection as reported by Watson and colleagues (2014). **A)** Schematic of the workflow for DNA library preparation and sequencing. **B)** Schematic of the workflow for bioinformatic analysis of the data produced. MAF = Minor allele frequency.

or if present at high frequencies in controls, and finally segregation testing performed on variants that look promising. This pipeline is typical of many NGS projects that have been reported in recent years, showing that most research groups within the field of genomics have encountered similar problems and developed similar solutions.

7.2.1 Further evidence of pathogenicity

Although some of the early publications in the literature reported variants as being pathogenic based solely on bioinformatic analysis and filtering methods, most groups have since come to the conclusion that further evidence of pathogenicity is required beyond simple validation and segregation analysis through Sanger sequencing. This is especially true of variants which have not previously been reported in the literature.

Chapter 5 of this thesis described experiments that showed the use of cDNA sequencing to provide further evidence to determine the effect of identified genomic mutations. These relatively simple experiments can quickly exclude the possibility of some variants being pathogenic, and can add greater support to the case for pathogenicity of others.

The sequencing of cDNA in these experiments were restricted to small specific targets, however the utility of mRNA to provide medically useful information has led to the growth of “RNA-seq” which involves using next-generation sequencing technologies on cDNA reverse transcribed from mRNA extracted from a specific tissue type to determine the sequence of all the transcripts in that tissue type.

RNA-seq has the potential to be a powerful tool, and when combined with whole genome sequencing of genomic DNA, the two data sets can be combined to determine the effects that mutations in the genome have on transcription. This study of RNA sequencing data has been dubbed “transcriptomics”, and it is likely that much future work will be directed to this field, and to how it can provide data to complement genomics and improve our understanding of genetic diseases.

One of the main disadvantages of using RNA-seq however, is the instability of RNA molecules, which can prevent long term storage of samples. In some cases this might mean samples from a patient are not available for testing if the patient is deceased, as was the case with patient OTH 12 (section 5.3.2.1 p. 183). In this case it was necessary to test samples from both parents instead to confirm a variant was causing a splicing effect.

7.2.2 Reference Genomes

In addition to the challenges presented in developing bioinformatic pipelines to analyse and filter variants, another problem can occur earlier in the processing of sequencing data, when aligning sequencing reads to a reference genome.

The reference sequence most commonly used for sequence alignment in NGS projects, including the work described in this thesis, is the human genome reference sequence (RefSeq) maintained by the Genome Reference Consortium (GRC, <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/>). The most recent RefSeq build released is GRCh38. The work in this project uses two earlier builds which were dubbed hg18 and hg19.

The GRC reference sequences are all relics of the original human genome project, and as such they were constructed to attempt to piece together a map of the human genome as quickly as possible and despite the limitations of the sequencing technology available at the time (Lander *et al.* 2001, Venter *et al.* 2001). As a result the RefSeq is known to be only approximately 99% complete, with many gaps mainly covering repetitive regions (Church *et al.* 2011).

The presence of inaccuracies in a reference sequence can have many implications on a sequencing project. Most sequence capture and enrichment methods use probes to capture target sequence, and these probes are designed using a reference sequence. If the reference sequence is incorrect, then the probes will not hybridise to the target region of the genome. As a result, genomic loci containing pathogenic variants would not be sequenced, with no indication of this on any of the sequencing metrics in the results.

The “Platinum genome” is a term that has been used to describe efforts to create more accurate reference sequences with fewer gaps, and to do this researchers have had to use sequencing methods that allow longer reads which can accurately map large repeat sequences, such as those found in telomeres, and other repeat-rich regions of the genome that have previously been mapped poorly in previous reference sequence builds. It is likely that at least some of these repetitive regions in the genome will harbour disease causing variants.

In addition to using longer reads, efforts to create a Platinum genome have used DNA from cell lines derived from hydatidiform moles, as shown in **Figure 7.3**. Hydatidiform moles are created when an egg that lacks maternal DNA is fertilised by a sperm. The paternal DNA is then replicated to create a diploid cell derived from just one parent, creating an entirely homozygous genome. When DNA from these cells is sequenced it is indistinguishable to sequencing DNA from a haploid genome. This avoids the computational challenges presented by having to sequence a diploid genome, and allows greater accuracy of the finished genome as a result.

The first genome of a hydatidiform mole was published by Steinberg and colleagues, however the authors admit that significant challenges remain to obtain a fully accurate reference sequence, as there are still large gaps in their published sequence (Steinberg *et al.* 2014).

TO SIMPLIFY A SEQUENCE

To produce uninterrupted DNA sequences of human chromosomes, geneticists are turning to hydatidiform moles. These are formed when a sperm cell enters an egg that has lost its nucleus, making it non-viable.

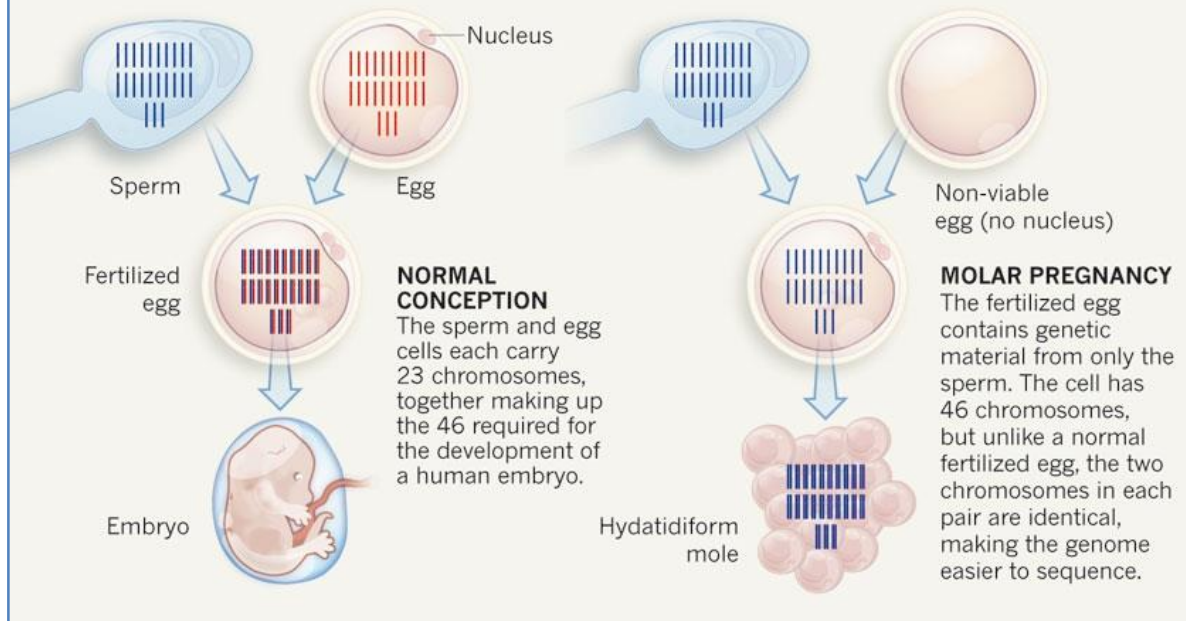


Figure 7.3 Creating a more accurate reference sequence using hydatidiform moles. Figure from: <http://www.nature.com/news/platinum-genome-takes-on-disease-1.16375>

7.3 Remaining Challenges

The work in Chapter 6 described experiments on samples from a cohort of patients with macular dystrophy in whom only a single recessive pathogenic mutation in the *ABCA4* gene had been identified, or none at all. The introduction in that chapter described a variety of potential explanations for why the missing pathogenic variants had not been identified (see p.209). Several of these hypotheses were tested, in particular the possible presence of cryptic exons in the *ABCA4* gene which are activated by mutations in intronic DNA.

During the course of this work a publication by Braun and colleagues proved that such variants exist, and in their cohort accounted for half the missing pathogenic variants (Braun *et al.* 2013). This still leaves approximately 10% of the total patients in their cohort with Stargardt disease without a secure molecular diagnosis.

It is probable that in some of these patients the disease is being caused by one of the other hypotheses described in the introduction to Chapter 6 (section 6.1.4 p. 209), for which there was not enough time to test. These explanations include mosaicism, large deletions, micro RNA binding, trinucleotide repeat expansion, other expansions, epigenetic effects, or errors with the sequencing methods so far used.

In the case of large deletions and errors in the sequencing methods used so far, the experiments using long range PCR amplification were attempted to provide a potential solution to both these problems. Amplification based methods of sequence capture and enrichment introduce far less bias to sequencing results compared to methods that rely on short tiled hybridisation probes to capture DNA. This should in theory remove the problem caused by an inaccurate reference sequence being used to design capture probes for hybridisation based capture methods.

Similarly, if a *de novo* assembly is performed on the data generated from the sequencing, it should also remove the problems caused by using an inaccurate reference sequence for the alignment of sequencing reads.

An alternative approach to identify whether the reference sequence is erroneous would be to run the data through an assembly algorithm and create a file of reads which do not map to the reference sequence. These reads which do not map would then have to be carefully analysed to determine whether they are in fact reads from the *ABCA4* gene which have not mapped due to a large deletion or other rearrangement in the gene.

This work will have to be performed in the future, as it will take time and will likely require the generation of new complex algorithms to analyse the data contained in individual sequence reads.

7.3.1 Third generation sequencing technologies

Over the past few years sequencing by synthesis using the Illumina methodology has become the dominant method of next-generation sequencing. New technologies that have been termed third generation are currently being developed which may present a higher throughput alternative to the Illumina chemistry, and which might also eliminate some of the weaknesses of the Illumina methodology.

One of the biggest disadvantages of Illumina's technology is the comparatively short read lengths, which makes it more difficult to detect large rearrangements and deletions. In some cases, these phenomena can only be detected by careful bioinformatic analysis of sequence reads which have been rejected by alignment software, which can easily be overlooked when dealing with such vast amounts of data.

Oxford Nanopore are a biotech company that have been developing a next-generation sequencing method that utilises very long sequence reads. The technology sequences single molecules of DNA with no prior amplification step. The DNA passes through a microscopic pore which can identify each nucleotide as it passes through by changes in the electrochemical charge of each base (Kasianowicz *et al.* 1996, Venkatesan and Bashir 2011).

Oxford Nanopore claim that when released their technology will be able to sequence reads of up to 10kbp, and that their library preparation process is minimal. The throughput of the technology is also expected to be able to provide high depth of coverage sequencing of a human genome for under \$1000, although no experimental data has yet been published proving this.

The ability to utilise such long reads would be of great benefit in detecting large rearrangements or deletions, as well as accurately mapping highly repetitive regions of the genome.

Another technology which produces very long reads is manufactured by Pacific Biosciences. The technology analyses single DNA molecules in real time by detecting the addition of individually fluorescently labelled nucleotides using zero-mode waveguide nanostructure arrays (Eid *et al.* 2009). The PacBio RS II sequencer can produce reads of up to 10-15kb in length.

7.4 The future of genomics

The 1000 genomes project was launched in 2008 and was the first project which aimed to sequence a large number of genomes. Its primary goal was to “create a complete and detailed catalogue of human genetic variations”. The pilot paper for the project was published on 28th October 2010, shortly after the commencement of the work described in this thesis (The Genomes Project 2010). For the majority of the duration of this project the 1000 genomes was the largest resource of human genomes available to use for analysing the frequency of variants identified compared to controls.

However, such has been the progress of reducing the cost of sequencing that on the 5th July 2013 the NHS in England announced the launch of the 100,000 genomes project which aims to sequence 100,000 genomes by the end of 2017. The project will be overseen by a new NHS owned and funded company called Genomics England.

The project will use samples from patients with rare diseases and cancer, recruited from NHS hospitals in London, Cambridge, Liverpool, Manchester, Oxford, Devon & Exeter, Southampton,

Birmingham, and Newcastle (<http://www.england.nhs.uk/2014/12/22/genomics-project/>). The project intends to sequence approximately 75,000 patient samples, with a further 25,000 genomes sequenced from relatives or tumour samples. Sequencing and analysis of data will both be performed by Illumina, with results then being sent back to the NHS for validation and clinical interpretation.

This is the first step in what many predict to be “the genomic revolution”, which will eventually lead to the sequencing of the genome of every member of the population. This could potentially create a vast resource of genetic data to be used by researchers to perform statistical analyses, identifying correlations between individual variants and the phenotypes of rare disease.

The information available to the individual would also be of great use. Hypothetically it would be possible to sequence the genomes of newborns, identify which diseases they are susceptible to, and begin planning treatments and gene therapies to pre-emptively treat the conditions before they cause damage.

A larger project has been announced in the USA, known as the precision medicine initiative, which aims to sequence one million genomes for a range of conditions including cancers, as well as rare diseases (Collins and Varmus 2015). This project has the long term aim of providing precision or personalised medicine: the ability to tailor drugs and therapies to an individual to provide the correct therapeutic dose. This could also allow the use of drugs that are shown to be effective but are not licenced by authorities such as the FDA due to the potential for side effects in a very small minority of patients with a specific genotype. Such drugs could hypothetically be licenced under a law that requires genetic testing to confirm it will not have a negative effect in the individual.

The potential promises of genomic technologies are great, but they also present ethical and societal implications. Some of the issues that have already arisen concern control over information, the ethics of incidental findings, and the patentability of genes.

The data obtained from whole genome sequencing could in theory inform the patient that they have a genetic mutation predisposing them to suffer from a neurodegenerative disease such as Huntington's or Alzheimer's later in life. These are incurable conditions, and many people given the choice, prefer not to have genetic testing performed on the genes associated with these conditions, or the results relayed to them in the case of incidental findings from whole genome sequencing.

Data from whole genome sequencing of the population could potentially radically alter the premiums demanded by insurance companies. Currently patients who have been diagnosed with existing conditions are required to declare them if they seek medical insurance, and the insurance company is entitled to ask for higher premiums, or for the patient to be given the offer of not being covered for their specific condition. The current plans for the 100,000 genomes project is for patient data to be anonymised when presented publicly to researchers, and other organisations, with only the patients' doctor allowed access to the de-anonymised results, so that clinical decisions can be made that are in the patients' interest.

One of the most high profile case of gene patenting so far has been the patents held by Myriad Genetics on the *BRCA1* and *BRCA2* genes. Mutations in these genes greatly increase the risk of ovarian and breast cancer in women. Patents filed by Myriad Genetics in 1994 and 1995 gave the company the sole right to develop and market diagnostic tests for these genes in the USA. This was met with a great deal of criticism and controversy from both patients and the general public in the USA. A lawsuit was successfully brought against the company arguing that the patents prevented patients from receiving a second opinion from another provider, which is a legal right under US law. In 2013 the US Supreme Court upheld the lawsuit, overturning the patents of *BRCA1* and *BRCA2* held by Myriad, and ruled that gene sequences cannot be patented as they are a product of nature (http://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf).

The vast changes created in society by the implications of genomic technologies, demonstrates the potential genomic medicine has to transform society and impact on human lives, and how it has already begun to do so.

7.5 Conclusion

Overall this project has shown that next-generation sequencing is capable of effectively and accurately interrogating large quantities of DNA to uncover the pathogenic variants that cause genetic disease.

The field of NGS and genomic medicine is growing rapidly, with improvements being made to all aspects of methodology, as well as data analysis. Significant challenges remain to be overcome, however the technology without doubt has the potential and the ability to change medical outcomes for patients.

This has been demonstrated in the work reported in this thesis, where the results of experiments using genomic technologies has produced medically useful information that has changed the prognosis and advice given to patients. This will also act as a first step in one day being able to provide gene therapy and other treatments that can prevent damage being caused by these conditions, and perhaps one day reverse it.

8. References

- Adachi, M., Y. Abe, Y. Aoki and Y. Matsubara (2012). "Epilepsy in RAS/MAPK syndrome: two cases of cardio-facio-cutaneous syndrome with epileptic encephalopathy and a literature review." Seizure **21**(1): 55-60.
- Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nat Meth **7**(4): 248-249.
- Akagi, T., M. Mandai, S. Ooto, Y. Hiram, F. Osakada, R. Kageyama, N. Yoshimura and M. Takahashi (2004). "Otx2 homeobox gene induces photoreceptor-specific phenotypes in cells derived from adult iris and ciliary tissue." Invest Ophthalmol Vis Sci **45**(12): 4570-4575.
- Albert, T. J., M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock and R. A. Gibbs (2007). "Direct selection of human genomic loci by microarray hybridization." Nat Methods **4**(11): 903-905.
- Allikmets, R. (2000). "Further Evidence for an Association of ABCR Alleles with Age-Related Macular Degeneration." The American Journal of Human Genetics **67**(2): 487-491.
- Allikmets, R., N. F. Shroyer, N. Singh, J. M. Seddon, R. A. Lewis, P. S. Bernstein, A. Peiffer, N. A. Zabriskie, Y. Li, A. Hutchinson, M. Dean, J. R. Lupski and M. Leppert (1997a). "Mutation of the Stargardt Disease Gene (ABCR) in Age-Related Macular Degeneration." Science **277**(5333): 1805-1807.
- Allikmets, R., N. Singh, H. Sun, N. F. Shroyer, A. Hutchinson, A. Chidambaram, B. Gerrard, L. Baird, D. Stauffer, A. Peiffer, A. Rattner, P. Smallwood, Y. Li, K. L. Anderson, R. A. Lewis, J. Nathans, M. Leppert, M. Dean and J. R. Lupski (1997b). "A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy." Nat Genet **15**(3): 236-246.
- Almeida, A., M. Layton and A. Karadimitris (2009). "Inherited glycosylphosphatidyl inositol deficiency: A treatable CDG." Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease **1792**(9): 874-880.
- Almeida, A. M., Y. Murakami, D. M. Layton, P. Hillmen, G. S. Sellick, Y. Maeda, S. Richards, S. Patterson, I. Kotsianidis, L. Mollica, D. H. Crawford, A. Baker, M. Ferguson, I. Roberts, R. Houlston, T. Kinoshita and A. Karadimitris (2006). "Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency." Nat Med **12**(7): 846-851.
- Anderson, D. H., M. J. Radeke, N. B. Gallo, E. A. Chapin, P. T. Johnson, C. R. Curletti, L. S. Hancox, J. Hu, J. N. Ebright, G. Malek, M. A. Hauser, C. B. Rickman, D. Bok, G. S. Hageman and L. V. Johnson (2010). "The pivotal role of the complement system in aging and age-related macular degeneration: hypothesis re-visited." Prog Retin Eye Res **29**(2): 95-112.
- Armstrong, J. D., D. Meyer, S. Xu and J. L. Elfervig (1998). "Long-term follow-up of Stargardt's disease and fundus flavimaculatus." Ophthalmology **105**(3): 448-457; discussion 457-448.
- Asan, Y. Xu, H. Jiang, C. Tyler-Smith, Y. Xue, T. Jiang, J. Wang, M. Wu, X. Liu, G. Tian, J. Wang, J. Wang, H. Yang and X. Zhang (2011). "Comprehensive comparison of three commercial human whole-exome capture platforms." Genome Biology **12**(9): R95.
- Audo, I., K. Bujakowska, T. Leveillard, S. Mohand-Said, M.-E. Lancelot, A. Germain, A. Antonio, C. Michiels, J.-P. Saraiva, M. Letexier, J.-A. Sahel, S. Bhattacharya and C. Zeitze (2012). "Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases." Orphanet Journal of Rare Diseases **7**(1): 8.
- Azarian, S. M. and G. H. Travis (1997). "The photoreceptor rim protein is an ABC transporter encoded by the gene for recessive Stargardt's disease (ABCR)." FEBS Lett **409**(2): 247-252.
- Bashir, R., A. Fatima and S. Naz (2010). "A frameshift mutation in SANS results in atypical Usher syndrome." Clin Genet **78**(6): 601-603.

- Baulac, S., G. Huberfeld, I. Gourfinkel-An, G. Mitropoulou, A. Beranger, J.-F. Prud'homme, M. Baulac, A. Brice, R. Bruzzone and E. LeGuern (2001). "First genetic evidence of GABAA receptor dysfunction in epilepsy: a mutation in the [gamma]2-subunit gene." Nat Genet **28**(1): 46-48.
- Beit-Ya'acov, A., L. Mizrahi-Meissonnier, A. Obolensky, C. Landau, A. Blumenfeld, A. Rosenmann, E. Banin and D. Sharon (2007). "Homozygosity for a novel ABCA4 founder splicing mutation is associated with progressive and severe Stargardt-like disease." Invest Ophthalmol Vis Sci **48**(9): 4308-4314.
- Belkadi, A., A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko, L. Shang, B. Boisson, J.-L. Casanova and L. Abel (2015). "Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants." Proceedings of the National Academy of Sciences **112**(17): 5473-5478.
- Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E. Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurler, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin and A. J. Smith (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.
- Bernstein, P. S., M. Leppert, N. Singh, M. Dean, R. A. Lewis, J. R. Lupski, R. Allikmets and J. M. Seddon (2002). "Genotype-phenotype analysis of ABCR variants in macular degeneration probands and siblings." Invest Ophthalmol Vis Sci **43**(2): 466-473.
- Biervert, C., B. C. Schroeder, C. Kubisch, S. F. Berkovic, P. Propping, T. J. Jentsch and O. K. Steinlein (1998). "A Potassium Channel Mutation in Neonatal Human Epilepsy." Science **279**(5349): 403-406.
- Blue, G. M., E. P. Kirk, E. Giannoulatou, S. L. Dunwoodie, J. W. K. Ho, D. C. K. Hilton, S. M. White, G. F. Sholler, R. P. Harvey and D. S. Winlaw (2014). "Targeted Next-Generation

- Sequencing Identifies Pathogenic Variants in Familial Congenital Heart Disease." Journal of the American College of Cardiology **64**(23): 2498-2506.
- Boughman, J. A., P. M. Conneally and W. E. Nance (1980). "Population genetic studies of retinitis pigmentosa." Am J Hum Genet **32**(2): 223-235.
- Bowne, S. J., S. P. Daiger, M. M. Hims, M. M. Sohocki, K. A. Malone, A. B. McKie, J. R. Heckenlively, D. G. Birch, C. F. Inglehearn, S. S. Bhattacharya, A. Bird and L. S. Sullivan (1999). "Mutations in the RP1 Gene Causing Autosomal Dominant Retinitis Pigmentosa." Human Molecular Genetics **8**(11): 2121-2128.
- Brancati, F., B. Dallapiccola and E. M. Valente (2010). "Joubert Syndrome and related disorders." Orphanet Journal of Rare Diseases **5**: 20-20.
- Braun, T. A., R. F. Mullins, A. H. Wagner, J. L. Andorf, R. M. Johnston, B. B. Bakall, A. P. Deluca, G. A. Fishman, B. L. Lam, R. G. Weleber, A. V. Cideciyan, S. G. Jacobson, V. C. Sheffield, B. A. Tucker and E. M. Stone (2013). "Non-exonic and synonymous variants in ABCA4 are an important cause of Stargardt disease." Human Molecular Genetics **22**(25): 5136-5145.
- Brockman, W., P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ, E. S. Lander, C. Nusbaum and D. B. Jaffe (2008). "Quality scores and SNP detection in sequencing-by-synthesis systems." Genome Research **18**(5): 763-770.
- Brouzes, E., M. Medkova, N. Savenelli, D. Marran, M. Twardowski, J. B. Hutchison, J. M. Rothberg, D. R. Link, N. Perrimon and M. L. Samuels (2009). "Droplet microfluidic technology for single-cell high-throughput screening." Proceedings of the National Academy of Sciences **106**(34): 14195-14200.
- Campuzano, V., L. Montermini, M. D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, F. Zara, J. Canizares, H. Koutnikova, S. I. Bidichandani, C. Gellera, A. Brice, P. Trouillas, G. De Michele, A. Filla, R. De Frutos, F. Palau, P. I. Patel, S. Di Donato, J. L. Mandel, S. Cocozza, M. Koenig and M. Pandolfo (1996). "Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion." Science **271**(5254): 1423-1427.
- Caraballo, R., S. Pavak, A. Lemaingue, M. Gastaldi, B. Echenne, J. Motte, P. Genton, R. Cersosimo, V. Humbertclaude, N. Fejerman, A. P. Monaco, M. G. Lathrop, J. Rochette and P. Szepietowski (2001). "Linkage of benign familial infantile convulsions to chromosome 16p12-q12 suggests allelism to the infantile convulsions and choreoathetosis syndrome." Am J Hum Genet **68**(3): 788-794.
- Carnevali, P., J. Baccash, A. L. Halpern, I. Nazarenko, G. B. Nilsen, K. P. Pant, J. C. Ebert, A. Brownley, M. Morenzoni, V. Karpinchyk, B. Martin, D. G. Ballinger and R. Drmanac (2012). "Computational techniques for human genome resequencing using mated gapped reads." J Comput Biol **19**(3): 279-292.
- Cartegni, L. and A. R. Krainer (2002). "Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1." Nat Genet **30**(4): 377-384.
- Castagnet, P., T. Mavlyutov, Y. Cai, F. Zhong and P. Ferreira (2003). "RPGRIP1s with distinct neuronal localization and biochemical properties associate selectively with RanBP2 in amacrine neurons." Hum Mol Genet **12**(15): 1847-1863.
- Chang, Y. F., J. S. Imam and M. F. Wilkinson (2007). "The nonsense-mediated decay RNA surveillance pathway." Annu Rev Biochem **76**: 51-74.
- Chapple, J. P., A. J. Hardcastle, C. Grayson, L. A. Spackman, K. R. Willison and M. E. Cheetham (2000). "Mutations in the N-terminus of the X-linked retinitis pigmentosa protein RP2 interfere with the normal targeting of the protein to the plasma membrane." Hum Mol Genet **9**(13): 1919-1926.
- Charlier, C., N. A. Singh, S. G. Ryan, T. B. Lewis, B. E. Reus, R. J. Leach and M. Leppert (1998). "A pore mutation in a novel KQT-like potassium channel gene in an idiopathic epilepsy family." Nat Genet **18**(1): 53-55.
- Chau, K. Y., S. Chen, D. J. Zack and S. J. Ono (2000). "Functional domains of the cone-rod homeobox (CRX) transcription factor." J Biol Chem **275**(47): 37264-37270.

- Chou, L.-S., C.-S. J. Liu, B. Boese, X. Zhang and R. Mao (2010). "DNA Sequence Capture and Enrichment by Microarray Followed by Next-Generation Sequencing for Targeted Resequencing: Neurofibromatosis Type 1 Gene as a Model." Clinical Chemistry **56**(1): 62-72.
- Church, D. M., V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. S. Ritchie, D. Albracht, M. Kremitzki, S. Rock, H. Kotkiewicz, C. Kremitzki, A. Wollam, L. Trani, L. Fulton, R. Fulton, L. Matthews, S. Whitehead, W. Chow, J. Torrance, M. Dunn, G. Harden, G. Threadgold, J. Wood, J. Collins, P. Heath, G. Griffiths, S. Pelan, D. Grafham, E. E. Eichler, G. Weinstock, E. R. Mardis, R. K. Wilson, K. Howe, P. Flicek and T. Hubbard (2011). "Modernizing Reference Genome Assemblies." PLoS Biol **9**(7): e1001091.
- Collins, F. S. and H. Varmus (2015). "A New Initiative on Precision Medicine." New England Journal of Medicine **372**(9): 793-795.
- Cremers, F. P., D. J. van de Pol, M. van Driel, A. I. den Hollander, F. J. van Haren, N. V. Knoers, N. Tijmes, A. A. Bergen, K. Rohrschneider, A. Blankenagel, A. J. Pinckers, A. F. Deutman and C. B. Hoyng (1998). "Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR." Hum Mol Genet **7**(3): 355-362.
- Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young and R. Jaenisch (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." Proceedings of the National Academy of Sciences **107**(50): 21931-21936.
- Daiger, S. P., L. S. Sullivan and J. A. Rodriguez (1995). "Correlation of phenotype with genotype in inherited retinal degeneration." Behavioral and Brain Sciences **18**(03): 452-467.
- Davies, W. I., S. M. Downes, J. K. Fu, M. E. Shanks, R. R. Copley, S. Lise, S. C. Ramsden, G. C. Black, K. Gibson, R. G. Foster, M. W. Hankins and A. H. Nemeth (2012). "Next-generation sequencing in health-care delivery: lessons from the functional analysis of rhodopsin." Genet Med **14**(11): 891-899.
- den Hollander, A. I., J. Davis, S. D. van der Velde-Visser, M. N. Zonneveld, C. O. Pierrottet, R. K. Koenekoop, U. Kellner, L. I. van den Born, J. R. Heckenlively, C. B. Hoyng, P. A. Handford, R. Roepman and F. P. Cremers (2004). "CRB1 mutation spectrum in inherited retinal dystrophies." Hum Mutat **24**(5): 355-369.
- den Hollander, A. I., J. B. ten Brink, Y. J. de Kok, S. van Soest, L. I. van den Born, M. A. van Driel, D. J. van de Pol, A. M. Payne, S. S. Bhattacharya, U. Kellner, C. B. Hoyng, A. Westerveld, H. G. Brunner, E. M. Bleeker-Wagemakers, A. F. Deutman, J. R. Heckenlively, F. P. Cremers and A. A. Bergen (1999). "Mutations in a human homologue of Drosophila crumbs cause retinitis pigmentosa (RP12)." Nat Genet **23**(2): 217-221.
- Denayer, E. and E. Legius (2007). "What's new in the neuro-cardio-facial-cutaneous syndromes?" European Journal of Pediatrics **166**(11): 1091-1098.
- Desmet, F.-O., D. Hamroun, M. Lalande, G. Collod-Bérout, M. Claustres and C. Bérout (2009). "Human Splicing Finder: an online bioinformatics tool to predict splicing signals." Nucleic Acids Research.
- Dixon-Salazar, T. J., J. L. Silhavy, N. Udpa, J. Schroth, S. Bielas, A. E. Schaffer, J. Olvera, V. Bafna, M. S. Zaki, G. H. Abdel-Salam, L. A. Mansour, L. Selim, S. Abdel-Hadi, N. Marzouki, T. Ben-Omran, N. A. Al-Saana, F. M. Sonmez, F. Celep, M. Azam, K. J. Hill, A. Collazo, A. G. Fenstermaker, G. Novarino, N. Akizu, K. V. Garimella, C. Sougnez, C. Russ, S. B. Gabriel and J. G. Gleeson (2012). "Exome sequencing can improve diagnosis and alter patient management." Sci Transl Med **4**(138): 138ra178.
- Dobrowolski, S. F., H. S. Andersen, T. K. Doktor and B. S. Andresen (2010). "The phenylalanine hydroxylase c.30C>G synonymous variation (p.G10G) creates a common exonic splicing silencer." Molecular Genetics and Metabolism **100**(4): 316-323.

- Donders, F. C. Z. H. (1884). Brieven van Franciscus Cornelis Donders (1818-1889) aan Hendrik Zwaardemaker (1857-1930) Coll. Zwa. 29 A 1-8. Coll. Zwa. 29 A 1-8.
- Downes, S. M., E. Packham, T. Cranston, P. Clouston, A. Seller and A. H. Nemeth (2012). "Detection rate of pathogenic mutations in *abca4* using direct sequencing: Clinical and research implications." Archives of Ophthalmology **130**(11): 1486-1490.
- Drmanac, R., S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zeremski, J. Snoddy, W. Funkhouser, B. Koop, L. Hood and a. et (1993). "DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing." Science **260**(5114): 1649-1652.
- Drmanac, R., A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcharding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church and C. A. Reid (2010). "Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays." Science **327**(5961): 78-81.
- Dryja, T. P., S. M. Adams, J. L. Grimsby, T. L. McGee, D. H. Hong, T. Li, S. Andreasson and E. L. Berson (2001). "Null RPGRIP1 alleles in patients with Leber congenital amaurosis." Am J Hum Genet **68**(5): 1295-1298.
- Dryja, T. P., J. A. McEvoy, T. L. McGee and E. L. Berson (2000). "Novel Rhodopsin Mutations Gly114Val and Gln184Pro in Dominant Retinitis Pigmentosa." Investigative Ophthalmology & Visual Science **41**(10): 3124-3127.
- Dryja, T. P., T. L. McGee, E. Reichel, L. B. Hahn, G. S. Cowley, D. W. Yandell, M. A. Sandberg and E. L. Berson (1990). "A point mutation of the rhodopsin gene in one form of retinitis pigmentosa." Nature **343**(6256): 364-366.
- Dvir, L., G. Srour, R. Abu-Ras, B. Miller, S. A. Shalev and T. Ben-Yosef (2010). "Autosomal-recessive early-onset retinitis pigmentosa caused by a mutation in PDE6G, the gene encoding the gamma subunit of rod cGMP phosphodiesterase." Am J Hum Genet **87**(2): 258-264.
- Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and S. Turner (2009). "Real-Time DNA Sequencing from Single Polymerase Molecules." Science **323**(5910): 133-138.
- Escayg, A., B. T. MacDonald, M. H. Meisler, S. Baulac, G. Huberfeld, I. An-Gourfinkel, A. Brice, E. LeGuern, B. Moulard, D. Chaigne, C. Buresi and A. Malafosse (2000). "Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2." Nat Genet **24**(4): 343-345.
- Eulalio, A., E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser and E. Izaurralde (2009). "Deadenylation is a widespread effect of miRNA regulation." Rna **15**(1): 21-32.
- Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-185.
- Fairbrother, W. G., R. F. Yeh, P. A. Sharp and C. B. Burge (2002). "Predictive identification of exonic splicing enhancers in human genes." Science **297**(5583): 1007-1013.
- Farkas, M., G. Grant, J. White, M. Sousa, M. Consugar and E. Pierce (2013). "Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of

- novel transcribed sequence via significant alternative splicing and novel genes." BMC Genomics **14**(1): 486.
- Fishman, G. A., E. M. Stone, D. A. Eliason, C. M. Taylor, M. Lindeman and D. J. Derlacki (2003). "ABCA4 gene sequence variations in patients with autosomal recessive cone-rod dystrophy." Arch Ophthalmol **121**(6): 851-855.
- Franceschetti, A. and J. Francois (1965). "[Fundus flavimaculatus]." Arch Ophtalmol Rev Gen Ophtalmol **25**(6): 505-530.
- Franke, R. R., B. Konig, T. P. Sakmar, H. G. Khorana and K. P. Hofmann (1990). "Rhodopsin mutants that bind but fail to activate transducin." Science **250**(4977): 123-125.
- Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." Proceedings of the National Academy of Sciences of the United States of America **89**(5): 1827-1831.
- Gerber, S., J. M. Rozet, D. Bonneau, E. Souied, A. Camuzat, J. L. Dufier, P. Amalric, J. Weissenbach, A. Munnich and J. Kaplan (1995). "A gene for late-onset fundus flavimaculatus with macular dystrophy maps to chromosome 1p13." Am J Hum Genet **56**(2): 396-399.
- Gerber, S., J. M. Rozet, T. J. R. van de Pol, C. B. Hoyng, A. Munnich, A. Blankenagel, J. Kaplan and F. P. M. Cremers (1998). "Complete Exon-Intron Structure of the Retina-Specific ATP Binding Transporter Gene (ABCR) Allows the Identification of Novel Mutations Underlying Stargardt Disease." Genomics **48**(1): 139-142.
- Gerth, C., M. Andrassi-Darida, M. Bock, M. N. Preising, B. H. Weber and B. Lorenz (2002). "Phenotypes of 16 Stargardt macular dystrophy/fundus flavimaculatus patients with known ABCA4 mutations and evaluation of genotype-phenotype correlation." Graefes Arch Clin Exp Ophthalmol **240**(8): 628-638.
- Glockle, N., S. Kohl, J. Mohr, T. Scheurenbrand, A. Sprecher, N. Weisschuh, A. Bernd, G. Rudolph, M. Schubach, C. Poloschek, E. Zrenner, S. Biskup, W. Berger, B. Wissinger and J. Neidhardt (2014). "Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies." Eur J Hum Genet **22**(1): 99-104.
- Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander and C. Nusbaum (2009). "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing." Nat Biotechnol **27**(2): 182-189.
- Gonzalez-Perez, A. and N. Lopez-Bigas (2011). "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel." Am J Hum Genet **88**(4): 440-449.
- Grant, G. R., M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch and E. A. Pierce (2011). "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)." Bioinformatics **27**(18): 2518-2528.
- Grawe, F., A. Wodarz, B. Lee, E. Knust and H. Skaer (1996). "The Drosophila genes crumbs and stardust are involved in the biogenesis of adherens junctions." Development **122**(3): 951-959.
- Gu, S.-m., D. A. Thompson, C. R. S. Srikumari, B. Lorenz, U. Finckh, A. Nicoletti, K. R. Murthy, M. Rathmann, G. Kumaramanickavel, M. J. Denton and A. Gal (1997). "Mutations in RPE65 cause autosomal recessive childhood-onset severe retinal dystrophy." Nat Genet **17**(2): 194-197.
- Guipponi, M., F. Rivier, F. Vigeveno, C. Beck, A. Crespel, B. Echenne, P. Lucchini, R. Sebastianelli, M. Baldy-Moulinier and A. Malafosse (1997). "Linkage mapping of benign familial infantile convulsions (BFIC) to chromosome 19q." Hum Mol Genet **6**(3): 473-477.
- Hameed, A., A. Abid, A. Aziz, M. Ismail, S. Q. Mehdi and S. Khaliq (2003). "Evidence of RPGRIP1 gene mutations associated with recessive cone-rod dystrophy." J Med Genet **40**(8): 616-619.

- Harris, J. K., J. W. Sahl, T. A. Castoe, B. D. Wagner, D. D. Pollock and J. R. Spear (2010). "Comparison of Normalization Methods for Construction of Large, Multiplex Amplicon Pools for Next-Generation Sequencing." Applied and Environmental Microbiology **76**(12): 3863-3868.
- Heron, S. E., K. M. Crossland, E. Andermann, H. A. Phillips, A. J. Hall, A. Bleasel, M. Shevell, S. Mercho, M.-H. Seni, M.-C. Guiot, J. C. Mulley, S. F. Berkovic and I. E. Scheffer (2002). "Sodium-channel defects in benign familial neonatal-infantile seizures." The Lancet **360**(9336): 851-852.
- Hoppman-Chaney, N., L. M. Peterson, E. W. Klee, S. Middha, L. K. Courteau and M. J. Ferber (2010). "Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics." Clin Chem **56**(8): 1297-1306.
- Hoyng, C. B., F. Poppelaars, T. J. van de Pol, H. Kremer, A. J. Pinckers, A. F. Deutman and F. P. Cremers (1996). "Genetic fine mapping of the gene for recessive Stargardt disease." Hum Genet **98**(4): 500-504.
- Illing, M., L. L. Molday and R. S. Molday (1997). "The 220-kDa Rim Protein of Retinal Rod Outer Segments Is a Member of the ABC Transporter Superfamily." Journal of Biological Chemistry **272**(15): 10303-10310.
- Inglehearn, C. F., T. J. Keen, R. Bashir, M. Jay, F. Fitzke, A. C. Bird, A. Crombie and S. Bhattacharya (1992). "A completed screen for mutations of the rhodopsin gene in a panel of patients with autosomal dominant retinitis pigmentosa." Hum Mol Genet **1**(1): 41-45.
- Ingram, V. M. (1956). "A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin." Nature **178**(4537): 792-794.
- Ionasescu, V. V., C. Searby, R. Ionasescu, I. M. Neuhaus and R. Werner (1996). "Mutations of the noncoding region of the connexin32 gene in X-linked dominant Charcot-Marie-Tooth neuropathy." Neurology **47**(2): 541-544.
- Jaakson, K., J. Zernant, M. Kulm, A. Hutchinson, N. Tonisson, D. Glavac, M. Ravnik-Glavac, M. Hawlina, M. R. Meltzer, R. C. Caruso, F. Testa, A. Maugeri, C. B. Hoyng, P. Gouras, F. Simonelli, R. A. Lewis, J. R. Lupski, F. P. Cremers and R. Allikmets (2003). "Genotyping microarray (gene chip) for the ABCR (ABCA4) gene." Hum Mutat **22**(5): 395-403.
- Johnston, Jennifer J., Andrea L. Gropman, Julie C. Sapp, Jamie K. Teer, Jodie M. Martin, Cyndi F. Liu, X. Yuan, Z. Ye, L. Cheng, Robert A. Brodsky and Leslie G. Biesecker (2012). "The Phenotype of a Germline Mutation in PIGA: The Gene Somatic Mutated in Paroxysmal Nocturnal Hemoglobinuria." The American Journal of Human Genetics **90**(2): 295-300.
- Kallioniemi, A., O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman and D. Pinkel (1992). "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors." Science **258**(5083): 818-821.
- Kaplan, J., S. Gerber, D. Larget-Piet, J. M. Rozet, H. Dollfus, J. L. Dufier, S. Odent, A. Postel-Vinay, N. Janin, M. L. Briard and et al. (1993). "A gene for Stargardt's disease (fundus flavimaculatus) maps to the short arm of chromosome 1." Nat Genet **5**(3): 308-311.
- Kasianowicz, J. J., E. Brandin, D. Branton and D. W. Deamer (1996). "Characterization of individual polynucleotide molecules using a membrane channel." Proceedings of the National Academy of Sciences **93**(24): 13770-13773.
- Kato, M., S. Das, K. Petras, K. Kitamura, K.-i. Morohashi, D. N. Abuelo, M. Barr, D. Bonneau, A. F. Brady, N. J. Carpenter, K. L. Ciperio, F. Frisone, T. Fukuda, R. Guerrini, E. Iida, M. Itoh, A. F. Lewanda, Y. Nanba, A. Oka, V. K. Proud, P. Saugier-veber, S. L. Schelley, A. Selicorni, R. Shaner, M. Silengo, F. Stewart, N. Sugiyama, J. Toyama, A. Toutain, A. L. Vargas, M. Yanazawa, E. H. Zackai and W. B. Dobyns (2004). "Mutations of ARX are associated with striking pleiotropy and consistent genotype-phenotype correlation." Human Mutation **23**(2): 147-159.
- Kato, M., S. Saitoh, A. Kamei, H. Shiraishi, Y. Ueda, M. Akasaka, J. Tohyama, N. Akasaka and K. Hayasaka (2007). "A Longer Polyalanine Expansion Mutation in the ARX Gene

- Causes Early Infantile Epileptic Encephalopathy with Suppression-Burst Pattern (Ohtahara Syndrome)." The American Journal of Human Genetics **81**(2): 361-366.
- Kelsell, R. E., K. Gregory-Evans, A. M. Payne, I. Perrault, J. Kaplan, R.-B. Yang, D. L. Garbers, A. C. Bird, A. T. Moore and D. M. Hunt (1998). "Mutations in the Retinal Guanylate Cyclase (RETGC-1) Gene in Dominant Cone-Rod Dystrophy." Human Molecular Genetics **7**(7): 1179-1184.
- Kent, W. J. (2002). "BLAT—The BLAST-Like Alignment Tool." Genome Research **12**(4): 656-664.
- Khaliq, S., A. Abid, M. Ismail, A. Hameed, A. Mohyuddin, P. Lall, A. Aziz, K. Anwar and S. Q. Mehdi (2005). "Novel association of RP1 gene mutations with autosomal recessive retinitis pigmentosa." J Med Genet **42**(5): 436-438.
- Khramtsov, N. V., E. A. Feshchenko, V. A. Suslova, B. E. Shmukler, B. E. Terpugov, T. V. Rakitina, N. V. Atabekova and V. M. Lipkin (1993). "The human rod photoreceptor cGMP phosphodiesterase beta-subunit. Structural studies of its cDNA and gene." FEBS Lett **327**(3): 275-278.
- Kitiratschky, V. B. D., D. Nagy, T. Zabel, E. Zrenner, B. Wissinger, S. Kohl and H. Jägle (2008). "Cone and cone-rod dystrophy segregating in the same pedigree due to the same novel CRX gene mutation." British Journal of Ophthalmology **92**(8): 1086-1091.
- Klevering, B. J., A. F. Deutman, A. Maugeri, F. P. M. Cremers and C. B. Hoyng (2005). "The spectrum of retinal phenotypes caused by mutations in the ABCR gene." Graefe's Archive for Clinical and Experimental Ophthalmology **243**(2): 90-100.
- Klevering, B. J., M. van Driel, D. J. van de Pol, A. J. Pinckers, F. P. Cremers and C. B. Hoyng (1999). "Phenotypic variations in a family with retinal dystrophy as result of different mutations in the ABCR gene." Br J Ophthalmol **83**(8): 914-918.
- Koressaar, T. and M. Remm (2007). "Enhancements and modifications of primer design program Primer3." Bioinformatics **23**(10): 1289-1291.
- Krawitz, P. M., M. R. Schweiger, C. Rodelsperger, C. Marcelis, U. Kolsch, C. Meisel, F. Stephani, T. Kinoshita, Y. Murakami, S. Bauer, M. Isau, A. Fischer, A. Dahl, M. Kerick, J. Hecht, S. Kohler, M. Jager, J. Grunhagen, B. J. de Condor, S. Doelken, H. G. Brunner, P. Meinecke, E. Passarge, M. D. Thompson, D. E. Cole, D. Horn, T. Roscioli, S. Mundlos and P. N. Robinson (2010). "Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome." Nat Genet **42**(10): 827-829.
- Kumar, P., S. Henikoff and P. C. Ng (2009). "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." Nat. Protocols **4**(8): 1073-1081.
- Kurg, A., N. Tonisson, I. Georgiou, J. Shumaker, J. Tollett and A. Metspalu (2000). "Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology." Genet Test **4**(1): 1-7.
- Lamb, T. D. and E. N. Pugh, Jr. (2006). "Phototransduction, dark adaptation, and rhodopsin regeneration the proctor lecture." Invest Ophthalmol Vis Sci **47**(12): 5137-5152.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczký, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chisoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M.

- Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). "Initial sequencing and analysis of the human genome." *Nature* **409**(6822): 860-921.
- Leamon, J. H., W. L. Lee, K. R. Tartaro, J. R. Lanza, G. J. Sarkis, A. D. deWinter, J. Berka and K. L. Lohman (2003). "A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions." *ELECTROPHORESIS* **24**(21): 3769-3777.
- Lelieveld, S. H., M. Spielmann, S. Mundlos, J. A. Veltman and C. Gilissen (2015). "Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions." *Hum Mutat* **36**(8): 815-822.
- Lewis, R. A., N. F. Shroyer, N. Singh, R. Allikmets, A. Hutchinson, Y. Li, J. R. Lupski, M. Leppert and M. Dean (1999). "Genotype/Phenotype analysis of a photoreceptor-specific ATP-binding cassette transporter gene, ABCR, in Stargardt disease." *Am. J. Hum. Genet.* **64**(2): 422-434.
- Li, J. Z., B. Chapman, P. Charlebois, O. Hofmann, B. Weiner, A. J. Porter, R. Samuel, S. Vardhanabhuti, L. Zheng, J. Eron, B. Taiwo, M. C. Zody, M. R. Henn, D. R. Kuritzkes, W. Hide, C. C. Wilson, B. I. Berzins, E. P. Acosta, B. Bastow, P. S. Kim, S. W. Read, J. Janik, D. S. Meres, M. M. Lederman, L. Mong-Kryspin, K. E. Shaw, L. G. Zimmerman, R. Leavitt, G. De La Rosa and A. Jennings (2014). "Comparison of illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy." *PLoS One* **9**(3): e90485.
- Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012). "Comparison of next-generation sequencing systems." *J Biomed Biotechnol* **2012**: 251364.
- Liu, M. M., E. Agrón, E. Chew, C. Meyerle, F. L. Ferris, C.-C. Chan and J. Tuo (2011). "Copy Number Variations in Candidate Genes in Neovascular Age-Related Macular Degeneration." *Investigative Ophthalmology & Visual Science* **52**(6): 3129-3135.
- Liu, Q., J. Zhou, S. P. Daiger, D. B. Farber, J. R. Heckenlively, J. E. Smith, L. S. Sullivan, J. Zuo, A. H. Milam and E. A. Pierce (2002). "Identification and subcellular localization of the RP1 protein in human and mouse photoreceptors." *Invest Ophthalmol Vis Sci* **43**(1): 22-32.
- Ludlow, L. B., B. P. Schick, M. L. Budarf, D. A. Driscoll, E. H. Zackai, A. Cohen and B. A. Konkle (1996). "Identification of a mutation in a GATA binding site of the platelet

- glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome." J Biol Chem **271**(36): 22076-22080.
- Lunter, G. and M. Goodson (2011). "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads." Genome Research **21**(6): 936-939.
- Marcinkowska-Swojak, M., B. Uszczynska, M. Figlerowicz and P. Kozlowski (2013). "An MLPA-based strategy for discrete CNV genotyping: CNV-miRNAs as an example." Hum Mutat **34**(5): 763-773.
- Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
- Marlhens, F., C. Bareil, J.-M. Griffoin, E. Zrenner, P. Amalric, C. Eliaou, S.-Y. Liu, E. Harris, T. M. Redmond, B. Arnaud, M. Claustres and C. P. Hamel (1997). "Mutations in RPE65 cause Leber's congenital amaurosis." Nat Genet **17**(2): 139-141.
- Martin, H. C., G. E. Kim, A. T. Pagnamenta, Y. Murakami, G. L. Carvill, E. Meyer, R. R. Copley, A. Rimmer, G. Barcia, M. R. Fleming, J. Kronengold, M. R. Brown, K. A. Hudspith, J. Broxholme, A. Kanapin, J. B. Cazier, T. Kinoshita, R. Nabbout, D. Bentley, G. McVean, S. Heavin, Z. Zaiwalla, T. McShane, H. C. Mefford, D. Shears, H. Stewart, M. A. Kurian, I. E. Scheffer, E. Blair, P. Donnelly, L. K. Kaczmarek and J. C. Taylor (2014). "Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis." Hum Mol Genet **23**(12): 3200-3211.
- Martinez-Mir, A., E. Paloma, R. Allikmets, C. Ayuso, T. d. Rio, M. Dean, L. Vilageliu, R. Gonzalez-Duarte and S. Balcells (1998). "Retinitis pigmentosa caused by a homozygous mutation in the Stargardt disease gene ABCR." Nat Genet **18**(1): 11-12.
- Maskos, U. and E. M. Southern (1992). "Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ." Nucleic Acids Research **20**(7): 1679-1684.
- Maston, G. A., S. K. Evans and M. R. Green (2006). "Transcriptional regulatory elements in the human genome." Annu Rev Genomics Hum Genet **7**: 29-59.
- Matsumoto, H. and C. A. Marsan (1964). "Cellular Mechanisms in Experimental Epileptic Seizures." Science **144**(3615): 193-194.
- Maugeri, A., B. J. Klevering, K. Rohrschneider, A. Blankenagel, H. G. Brunner, A. F. Deutman, C. B. Hoyng and F. P. M. Cremers (2000). "Mutations in the ABCA4 (ABCR) Gene Are the Major Cause of Autosomal Recessive Cone-Rod Dystrophy." American journal of human genetics **67**(4): 960-966.
- Maugeri, A., M. A. van Driel, D. J. R. van de Pol, B. J. Klevering, F. J. J. van Haren, N. Tijmes, A. A. B. Bergen, K. Rohrschneider, A. Blankenagel, A. J. L. G. Pinckers, N. Dahl, H. G. Brunner, A. F. Deutman, C. B. Hoyng and F. P. M. Cremers (1999). "The 2588G'C Mutation in the ABCR Gene Is a Mild Frequent Founder Mutation in the Western European Population and Allows the Classification of ABCR Mutations in Patients with Stargardt Disease." American journal of human genetics **64**(4): 1024-1035.
- Maydan, G., I. Noyman, A. Har-Zahav, Z. B. Neria, M. Pasmanik-Chor, A. Yeheskel, A. Albin-Kaplanski, I. Maya, N. Magal, E. Birk, A. J. Simon, A. Halevy, G. Rechavi, M. Shohat, R. Straussberg and L. Basel-Vanagaite (2011). "Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN." Journal of Medical Genetics **48**(6): 383-389.
- McBean, R. S., C. A. Hyland, J. L. Hendry, M.-T. Shabani-Rad and R. L. Flower (2015). "SARA: a "new" low-frequency MNS antigen (MNS47) provides further evidence of the extreme diversity of the MNS blood group system." Transfusion: n/a-n/a.

- McLaughlin, M. E., M. A. Sandberg, E. L. Berson and T. P. Dryja (1993). "Recessive mutations in the gene encoding the beta-subunit of rod phosphodiesterase in patients with retinitis pigmentosa." Nat Genet **4**(2): 130-134.
- Menon, S. T., M. Han and T. P. Sakmar (2001). "Rhodopsin: structural basis of molecular physiology." Physiol Rev **81**(4): 1659-1688.
- Mi, H., A. Muruganujan, J. T. Casagrande and P. D. Thomas (2013). "Large-scale gene function analysis with the PANTHER classification system." Nat. Protocols **8**(8): 1551-1566.
- Mitton, K. P., P. K. Swain, S. Chen, S. Xu, D. J. Zack and A. Swaroop (2000). "The Leucine Zipper of NRL Interacts with the CRX Homeodomain." Journal of Biological Chemistry **275**(38): 29794-29799.
- Moiseyev, G., Y. Chen, Y. Takahashi, B. X. Wu and J. X. Ma (2005). "RPE65 is the isomerohydrolase in the retinoid visual cycle." Proc Natl Acad Sci U S A **102**(35): 12413-12418.
- Molday, L. L., A. R. Rabin and R. S. Molday (2000). "ABCR expression in foveal cone photoreceptors and its role in Stargardt macular dystrophy." American Journal of Ophthalmology **130**(5): 689.
- Morgan, J. E., I. M. Carr, E. Sheridan, C. E. Chu, B. Hayward, N. Camm, H. A. Lindsay, C. J. Mattocks, A. F. Markham, D. T. Bonthron and G. R. Taylor (2010). "Genetic diagnosis of familial breast cancer using clonal sequencing." Human Mutation **31**(4): 484-491.
- Morgan, T. H. (1911). "RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE." Science **34**(873): 384.
- Mullis, K. B. and F. A. Faloona (1987). "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction." Methods Enzymol **155**: 335-350.
- Neveling, K., R. W. J. Collin, C. Gilissen, R. A. C. van Huet, L. Visser, M. P. Kwint, S. J. Gijzen, M. N. Zonneveld, N. Wieskamp, J. de Ligt, A. M. Siemiatkowska, L. H. Hoefsloot, M. F. Buckley, U. Kellner, K. E. Branham, A. I. den Hollander, A. Hoischen, C. Hoyng, B. J. Klevering, L. I. van den Born, J. A. Veltman, F. P. M. Cremers and H. Scheffer (2012). "Next-generation genetic testing for retinitis pigmentosa." Human Mutation: n/a-n/a.
- Nyren, P. and V. Edwin (1994). "Inorganic Pyrophosphatase-Based Detection Systems: II. Detection and Quantification of Cell Lysis and Cell-Lysing Activity." Analytical Biochemistry **220**(1): 46-52.
- Nyren, P. and A. Lundin (1985). "Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis." Anal Biochem **151**(2): 504-509.
- O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure and E. E. Eichler (2012). "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations." Nature **485**(7397): 246-250.
- O'Sullivan, J., B. G. Mullaney, S. S. Bhaskar, J. E. Dickerson, G. Hall, A. O'Grady, A. Webster, S. C. Ramsden and G. C. Black (2012). "A paradigm shift in the delivery of services for diagnosis of inherited retinal disease." J Med Genet **49**(5): 322-326.
- Ohtahara, S., Y. Ohtsuka and E. Oka (1997). "Epileptic encephalopathies in early infancy." Indian J Pediatr **64**(5): 603-612.
- Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi and T. Sekiya (1989). "Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms." Proc Natl Acad Sci U S A **86**(8): 2766-2770.
- Pagnamenta, A. T., S. Lise, V. Harrison, H. Stewart, S. Jayawant, G. Quaghebeur, A. T. Deng, V. E. Murphy, E. S. Akha, A. Rimmer, I. Mathieson, S. J. L. Knight, U. Kini, J. C. Taylor and D. A. Keays (2012). "Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies." J Hum Genet **57**(1): 70-72.
- Perez, B., F. Mechinaud, C. Galambrun, N. Ben Romdhane, B. Isidor, N. Philip, J. Derain-Court, B. Cassinat, J. Lachenaud, S. Kaltenbach, A. Salmon, C. Desiree, S. Pereira, M. L. Menot, N. Royer, O. Fenneteau, A. Baruchel, C. Chomienne, A. Verloes and H.

- Cave (2010). "Germline mutations of the CBL gene define a new genetic syndrome with predisposition to juvenile myelomonocytic leukaemia." *J Med Genet* **47**(10): 686-691.
- Pertea, M., X. Lin and S. L. Salzberg (2001). "GeneSplicer: a new computational method for splice site prediction." *Nucleic Acids Res* **29**(5): 1185-1190.
- Pierce, E. A., T. Quinn, T. Meehan, T. L. McGee, E. L. Berson and T. P. Dryja (1999). "Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa." *Nat Genet* **22**(3): 248-254.
- Prince, D. A. and B. J. Wilder (1967). "Control mechanisms in cortical epileptogenic foci. "Surround" inhibition." *Arch Neurol* **16**(2): 194-202.
- Quazi, F., S. Lenevich and R. S. Molday (2012). "ABCA4 is an N-retinylidene-phosphatidylethanolamine and phosphatidylethanolamine importer." *Nat Commun* **3**: 925.
- Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri and S. C. Schuster (2013). "Comparison of sequencing platforms for single nucleotide variant calls in a human sample." *PLoS One* **8**(2): e55089.
- Rett, A. and R. Teubel (1964). "Neugeborenenkrampe im Rahmen einer epileptisch belasteten Familie." *Wien Klin Wschr* **76**: 609-613.
- Reva, B., Y. Antipin and C. Sander (2011). "Predicting the functional impact of protein mutations: application to cancer genomics." *Nucleic Acids Res* **39**(17): e118.
- Richterich, P. (1998). "Estimation of errors in "raw" DNA sequences: a validation study." *Genome Res* **8**(3): 251-259.
- Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, W. G. S. Consortium, A. O. M. Wilkie, G. McVean and G. Lunter (2014). "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications." *Nat Genet* **46**(8): 912-918.
- Rivera, A., K. White, H. Stöhr, K. Steiner, N. Hemmrich, T. Grimm, B. Jurklies, B. Lorenz, H. P. N. Scholl, E. Apfelstedt-Sylla and B. H. F. Weber (2000). "A Comprehensive Survey of Sequence Variation in the ABCA4 (ABCR) Gene in Stargardt Disease and Age-Related Macular Degeneration." *The American Journal of Human Genetics* **67**(4): 800-813.
- Roberts, A., J. Allanson, S. K. Jadico, M. I. Kavamura, J. Noonan, J. M. Opitz, T. Young and G. Neri (2006). "The cardiofaciocutaneous syndrome." *J Med Genet* **43**(11): 833-842.
- Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). "Integrative genomics viewer." *Nat Biotech* **29**(1): 24-26.
- Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén and P. Nyrén (1996). "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release." *Analytical Biochemistry* **242**(1): 84-89.
- Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." *Nat Biotech* **26**(10): 1117-1124.
- Rozet, J.-M., S. Gerber, I. Ghazi, I. Perrault, D. Ducroq, E. Souied, A. Cabot, J.-L. Dufier, A. Munnich and J. Kaplan (1999). "Mutations of the retinal specific ATP binding transporter gene (ABCR) in a single family segregating both autosomal recessive retinitis pigmentosa RP19 and Stargardt disease: evidence of clinical heterogeneity at this locus." *Journal of Medical Genetics* **36**(6): 447-451.
- Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich and N. Arnheim (1985). "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia." *Science* **230**(4732): 1350-1354.
- Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." *Journal of Molecular Biology* **94**(3): 441-446.
- Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences* **74**(12): 5463-5467.
- Schouten, J. P., C. J. McElgunn, R. Waaijer, D. Zwiijnenburg, F. Diepvens and G. Pals (2002). "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification." *Nucleic Acids Res* **30**(12): e57.

- Schwahn, U., S. Lenzner, J. Dong, S. Feil, B. Hinzmann, G. van Duijnhoven, R. Kirschner, M. Hemberger, A. A. Bergen, T. Rosenberg, A. J. Pinckers, R. Fundele, A. Rosenthal, F. P. Cremers, H. H. Ropers and W. Berger (1998). "Positional cloning of the gene for X-linked retinitis pigmentosa 2." Nat Genet **19**(4): 327-332.
- Schwarz, J. M., C. Rodelsperger, M. Schuelke and D. Seelow (2010). "MutationTaster evaluates disease-causing potential of sequence alterations." Nat Meth **7**(8): 575-576.
- Senderek, J., J. S. Muller, M. Dusl, T. M. Strom, V. Guergueltcheva, I. Diepolder, S. H. Laval, S. Maxwell, J. Cossins, S. Krause, N. Muelas, J. J. Vilchez, J. Colomer, C. J. Mallebrera, A. Nascimento, S. Nafissi, A. Kariminejad, Y. Nilipour, B. Bozorgmehr, H. Najmabadi, C. Rodolico, J. P. Sieb, O. K. Steinlein, B. Schlotter, B. Schoser, J. Kirschner, R. Herrmann, T. Voit, A. Oldfors, C. Lindbergh, A. Urtizberea, M. von der Hagen, A. Hubner, J. Palace, K. Bushby, V. Straub, D. Beeson, A. Abicht and H. Lochmuller (2011). "Hexosamine biosynthetic pathway mutations cause neuromuscular transmission defect." Am J Hum Genet **88**(2): 162-172.
- September, A. V., A. A. Vorster, R. S. Ramesar and L. J. Greenberg (2004). "Mutation Spectrum and Founder Chromosomes for the ABCA4 Gene in South African Patients with Stargardt Disease." Investigative Ophthalmology & Visual Science **45**(6): 1705-1711.
- Shanks, M. E., S. M. Downes, R. R. Copley, S. Lise, J. Broxholme, K. A. Hudspith, A. Kwasniewska, W. I. Davies, M. W. Hankins, E. R. Packham, P. Clouston, A. Seller, A. O. Wilkie, J. C. Taylor, J. Ragoussis and A. H. Nemeth (2013). "Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease." Eur J Hum Genet **21**(3): 274-280.
- Shroyer, N. F., R. A. Lewis, R. Allikmets, N. Singh, M. Dean, M. Leppert and J. R. Lupski (1999). "The rod photoreceptor ATP-binding cassette transporter gene, ABCR, and retinal disease: from monogenic to multifactorial." Vision Research **39**(15): 2537-2544.
- Shroyer, N. F., R. A. Lewis and J. R. Lupski (2000). "Complex inheritance of ABCR mutations in Stargardt disease: linkage disequilibrium, complex alleles, and pseudodominance." Hum Genet **106**(2): 244-248.
- Shroyer, N. F., R. A. Lewis, A. N. Yatsenko, T. G. Wensel and J. R. Lupski (2001). "Cosegregation and functional analysis of mutant ABCR (ABCA4) alleles in families that manifest both Stargardt disease and age-related macular degeneration." Hum Mol Genet **10**(23): 2671-2678.
- Simpson, D. A., G. R. Clark, S. Alexander, G. Silvestri and C. E. Willoughby (2010). "Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa." Journal of Medical Genetics.
- Sims, D., I. Sudbery, N. E. Illott, A. Heger and C. P. Ponting (2014). "Sequencing depth and coverage: key considerations in genomic analyses." Nat Rev Genet **15**(2): 121-132.
- Singh, N. A., C. Charlier, D. Stauffer, B. R. DuPont, R. J. Leach, R. Melis, G. M. Ronen, I. Bjerre, T. Quattlebaum, J. V. Murphy, M. L. McHarg, D. Gagnon, T. O. Rosales, A. Peiffer, V. E. Anderson and M. Leppert (1998). "A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns." Nat Genet **18**(1): 25-29.
- Smith, P. J., C. Zhang, J. Wang, S. L. Chew, M. Q. Zhang and A. R. Krainer (2006). "An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers." Human Molecular Genetics **15**(16): 2490-2508.
- Sohrab, M. A., R. Allikmets, M. M. Guarnaccia and R. T. Smith (2010). "Preimplantation genetic diagnosis for stargardt disease." Am J Ophthalmol **149**(4): 651-655.e652.
- Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." Journal of Molecular Biology **98**(3): 503-517.
- Srour, M., J. Schwartzentruber, F. F. Hamdan, L. H. Ospina, L. Patry, D. Labuda, C. Massicotte, S. Dobrzeniecka, J. M. Capo-Chichi, S. Papillon-Cavanagh, M. E. Samuels, K. M. Boycott, M. I. Shevell, R. Laframboise, V. Desilets, B. Maranda, G. A. Rouleau, J. Majewski and J. L. Michaud (2012). "Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population." Am J Hum Genet **90**(4): 693-700.
- Staley, K. (2015). "Molecular mechanisms of epilepsy." Nat Neurosci **18**(3): 367-372.

- Stargardt, K. (1909). "Ueber familiare, progressive Degeneration in der Makulagegend des Auges." Graefes Arch. Ophthalmol. **71**(534).
- Steinberg, K. M., V. A. Schneider, T. A. Graves-Lindsay, R. S. Fulton, R. Agarwala, J. Huddleston, S. A. Shiryev, A. Morgulis, U. Surti, W. C. Warren, D. M. Church, E. E. Eichler and R. K. Wilson (2014). "Single haplotype assembly of the human genome from a hydatidiform mole." Genome Research **24**(12): 2066-2076.
- Stella, A., A. Wagner, K. Shito, S. M. Lipkin, P. Watson, G. Guanti, H. T. Lynch, R. Fodde and B. Liu (2001). "A Nonsense Mutation in MLH1 Causes Exon Skipping in Three Unrelated HNPCC Families." Cancer Research **61**(19): 7020-7024.
- Stenson, P. D., E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. Thomas, S. Abeyasinghe, M. Krawczak and D. N. Cooper (2003). "Human Gene Mutation Database (HGMD): 2003 update." Hum Mutat **21**(6): 577-581.
- Stone, E. A. and A. Sidow (2005). "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity." Genome Res **15**(7): 978-986.
- Strachan, T. and A. Read (2011). Human Molecular Genetics. New York, Garland Science.
- Sulonen, A.-M., P. Ellonen, H. Almus, M. Lepisto, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, H. Joensuu, T. Raivio, A. Suomalainen and J. Saarela (2011). "Comparison of solution-based exome capture methods for next generation sequencing." Genome Biology **12**(9): R94.
- Sun, H. and J. Nathans (1997). "Stargardt's ABCR is localized to the disc membrane of retinal rod outer segments." Nat Genet **17**(1): 15-16.
- Sun, H., P. M. Smallwood and J. Nathans (2000). "Biochemical defects in ABCR protein variants associated with human retinopathies." Nat Genet **26**(2): 242-246.
- Sung, C. H., C. M. Davenport and J. Nathans (1993). "Rhodopsin mutations responsible for autosomal dominant retinitis pigmentosa. Clustering of functional classes along the polypeptide chain." J Biol Chem **268**(35): 26645-26649.
- Swain, P. K., S. Chen, Q.-L. Wang, L. M. Affatigato, C. L. Coats, K. D. Brady, G. A. Fishman, S. G. Jacobson, A. Swaroop, E. Stone, P. A. Sieving and D. J. Zack (1997). "Mutations in the Cone-Rod Homeobox Gene Are Associated with the Cone-Rod Dystrophy Photoreceptor Degeneration." Neuron **19**(6): 1329-1336.
- Swaminathan, G. and A. Y. Tsygankov (2006). "The Cbl family proteins: Ring leaders in regulation of cell signaling." Journal of Cellular Physiology **209**(1): 21-43.
- Tepass, U. and E. Knust (1993). "Crumbs and Stardust Act in a Genetic Pathway That Controls the Organization of Epithelia in *Drosophila melanogaster*." Developmental Biology **159**(1): 311-326.
- The Genomes Project, C. (2010). "A map of human genome variation from population scale sequencing." Nature **467**(7319): 1061-1073.
- Tonisson, N., A. Kurg, K. Kaasik, E. Lohmussaar and A. Metspalu (2000). "Unravelling genetic data by arrayed primer extension." Clin Chem Lab Med **38**(2): 165-170.
- Tsujino, A., C. Maertens, K. Ohno, X.-M. Shen, T. Fukuda, C. M. Harper, S. C. Cannon and A. G. Engel (2003). "Myasthenic syndrome caused by mutation of the SCN4A sodium channel." Proceedings of the National Academy of Sciences **100**(12): 7377-7382.
- Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012). "Primer3--new capabilities and interfaces." Nucleic Acids Res **40**(15): e115.
- van der Burgt, I. (2007). "Noonan syndrome." Orphanet Journal of Rare Diseases **2**(1): 4.
- Venkatesan, B. M. and R. Bashir (2011). "Nanopore sensors for nucleic acid analysis." Nat Nano **6**(10): 615-624.
- Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L.

- Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The Sequence of the Human Genome." *Science* **291**(5507): 1304-1351.
- Vincent, A. L., J. Carroll, G. A. Fishman, A. Sauer, D. Sharp, P. Summerfelt, V. Williams, A. M. Dubis, S. Kohl and F. Wong (2013). "F45L Allele Does Not Cause Autosomal Dominant Retinitis Pigmentosa in a Large Caucasian Family." *Transl Vis Sci Technol* **2**(2): 4.
- Wallace, R. H., D. W. Wang, R. Singh, I. E. Scheffer, A. L. George, Jr., H. A. Phillips, K. Saar, A. Reis, E. W. Johnson, G. R. Sutherland, S. F. Berkovic and J. C. Mulley (1998). "Febrile seizures and generalized epilepsy associated with a mutation in the Na⁺-channel beta1 subunit gene SCN1B." *Nat Genet* **19**(4): 366-370.
- Walther, H., J. D. Lambert, R. S. Jones, U. Heinemann and B. Hamon (1986). "Epileptiform activity in combined slices of the hippocampus, subiculum and entorhinal cortex during perfusion with low magnesium medium." *Neurosci Lett* **69**(2): 156-161.
- Watson, C. M., M. El-Asrag, D. A. Parry, J. E. Morgan, C. V. Logan, I. M. Carr, E. Sheridan, R. Charlton, C. A. Johnson, G. Taylor, C. Toomes, M. McKibbin, C. F. Inglehearn and M. Ali (2014). "Mutation Screening of Retinal Dystrophy Patients by Targeted Capture from Tagged Pooled DNAs and Next Generation Sequencing." *PLoS ONE* **9**(8): e104281.
- Webb, T. R., D. A. Parfitt, J. C. Gardner, A. Martinez, D. Bevilacqua, A. E. Davidson, I. Zito, D. L. Thiselton, J. H. C. Ressa, M. Apergi, N. Schwarz, N. Kanuga, M. Michaelides, M. E. Cheetham, M. B. Gorin and A. J. Hardcastle (2012). "Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23)." *Human Molecular Genetics* **21**(16): 3647-3654.
- Weber, F., K. Fukino, M. Villalona-Calero and C. Eng (2005). "Limitations of Single-Strand Conformation Polymorphism Analysis As a High-Throughput Method for the Detection

- of EGFR Mutations in the Clinical Setting." *Journal of Clinical Oncology* **23**(24): 5847-5848.
- Webster, A. R., E. Håkonsen, A. J. Lotery, K. Vandenberg, T. L. Casavant, K. T. Oh, G. Beck, G. A. Fishman, B. L. Lam, A. Levin, J. R. Heckenlively, S. G. Jacobson, R. G. Weleber, V. C. Sheffield and E. M. Stone (2001). "An Analysis of Allelic Variation in the ABCA4 Gene." *Investigative Ophthalmology & Visual Science* **42**(6): 1179-1189.
- Wei, X., X. Ju, X. Yi, Q. Zhu, N. Qu, T. Liu, Y. Chen, H. Jiang, G. Yang, R. Zhen, Z. Lan, M. Qi, J. Wang, Y. Yang, Y. Chu, X. Li, Y. Guang and J. Huang (2011). "Identification of Sequence Variants in Genetic Disease-Causing Genes Using Targeted Next-Generation Sequencing." *PLoS ONE* **6**(12): e29500.
- Weigell-Weber, M., S. Fokstuen, B. Torok, G. Niemeyer, A. Schinzel and M. Hergersberg (2000). "Codons 837 and 838 in the retinal guanylate cyclase gene on chromosome 17p: hot spots for mutations in autosomal dominant cone-rod dystrophy?" *Arch Ophthalmol* **118**(2): 300.
- Weng, J., N. L. Mata, S. M. Azarian, R. T. Tzekov, D. G. Birch and G. H. Travis (1999). "Insights into the function of Rim protein in photoreceptors and etiology of Stargardt's disease from the phenotype in abcr knockout mice." *Cell* **98**(1): 13-23.
- West, W. (1841). "On a peculiar form of infantile convulsions." *Lancet* **1**: 724-725.
- Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X.-z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs and J. M. Rothberg (2008). "The complete genome of an individual by massively parallel DNA sequencing." *Nature* **452**(7189): 872-876.
- Wheless, J. W., P. A. Gibson, K. L. Rosbeck, M. Hardin, C. O'Dell, V. Whittemore and J. M. Pellock (2012). "Infantile spasms (West syndrome): update and resources for pediatricians and providers to share with parents." *BMC Pediatr* **12**: 108.
- White, M. B., M. Carvalho, D. Derse, S. J. O'Brien and M. Dean (1992). "Detecting single base substitutions as heteroduplex polymorphisms." *Genomics* **12**(2): 301-306.
- Williams, R., S. G. Peisajovich, O. J. Miller, S. Magdassi, D. S. Tawfik and A. D. Griffiths (2006). "Amplification of complex gene libraries by emulsion PCR." *Nat Meth* **3**(7): 545-550.
- Yang, Z., Y. Chen, C. Lillo, J. Chien, Z. Yu, M. Michaelides, M. Klein, K. A. Howes, Y. Li, Y. Kaminoh, H. Chen, C. Zhao, Y. Chen, Y. T. Al-Sheikh, G. Karan, D. Corbeil, P. Escher, S. Kamaya, C. Li, S. Johnson, J. M. Frederick, Y. Zhao, C. Wang, D. J. Cameron, W. B. Huttner, D. F. Schorderet, F. L. Munier, A. T. Moore, D. G. Birch, W. Baehr, D. M. Hunt, D. S. Williams and K. Zhang (2008). "Mutant prominin 1 found in patients with macular degeneration disrupts photoreceptor disk morphogenesis in mice." *J Clin Invest* **118**(8): 2908-2916.
- Yatsenko, A., N. Shroyer, R. Lewis and J. Lupski (2001). "Late-onset Stargardt disease is associated with missense mutations that map outside known functional regions of ABCR (ABCA4)." *Human Genetics* **108**(4): 346-355.
- Zhang, F., W. Gu, M. E. Hurles and J. R. Lupski (2009). "Copy number variation in human health, disease, and evolution." *Annu Rev Genomics Hum Genet* **10**: 451-481.
- Zhang, K., M. Kniazeva, M. Han, W. Li, Z. Yu, Z. Yang, Y. Li, M. L. Metzker, R. Allikmets, D. J. Zack, L. E. Kakuk, P. S. Lagali, P. W. Wong, I. M. MacDonald, P. A. Sieving, D. J. Figueroa, C. P. Austin, R. J. Gould, R. Ayyagari and K. Petrukhin (2001). "A 5-bp deletion in ELOVL4 is associated with two related forms of autosomal dominant macular dystrophy." *Nat Genet* **27**(1): 89-93.

9. Appendix

9.1 454 GS FLX sequencing methods

The following methods appendix describes the procedure by which next-generation sequencing using the NimbleGen sequence capture array and Roche 454 GS FLX sequencing machine with Titanium reagents were performed. These technologies were introduced in sections 1.6.1 and 1.8.1. A figure showing a brief overview of the separate stages of DNA library preparation, sequence capture, emulsion PCR, and sequencing was presented in Chapter 2 methods section 2.2.9.2 (**Figure 2.1**).

9.1.1 454 DNA library preparation

The DNA library preparation process was performed in a pre-PCR room to avoid contamination of the libraries with amplicons.

9.1.1.1 Dilution and quantification

15µl sample DNA was transferred to a fresh microcentrifuge tube and 35µl TE buffer added to create a 50µl dilution. The tubes were rotated on a benchtop rotator for 90 minutes and stored at 4°C.

Quantitation mixture:

199(n+3)µl of buffer was added to (n+3)µl of dye, where n = number of samples to be quantitated.

Buffer = Q32853 COMPONENTB: Quant-iT dsDNA BR buffer (Invitrogen)

199µl of this mixture was aliquoted into Qubit tubes, with 1µl of sample added to each tube. Two tubes were made up with 190µl of quantitation mixture and 10µl standard. The standards were used to calibrate a Qubit fluorometer with a BR DNA assay, before sample readings were taken as described in section 2.2.3.2 p.77.

9.1.1.2 Nebulisation

Volumes of DNA samples as determined by Qubit quantitation were made up to 100µl with TE buffer in a microcentrifuge tube. 500µl was added and the solution mixed with a pipette.

Nebuliser cups were assembled by affixing a nebuliser condenser tube around the aspiration tube, and then placing a stopping on the lid and pressing down firmly. DNA samples were loaded into the nebulisation cups.

Nebulisation tubing was attached to the gas inlet on the nebuliser cup inside a fume cupboard, and the other end of the tubing to a valve attached by tubing to a nitrogen gas tank. Approximately 30psi nitrogen gas was allowed to flow into the nebuliser cup for seventy seconds. The cups were spun down, and 2500µl Qiagen PBI added and mixed. The DNA/buffer mixture was transferred to QIAquick PCR purification columns and purified according to the manufacturer's protocol (see section 2.2.7.1 p.86). Each sample had its elutions pooled into a single microcentrifuge tube for that sample.

9.1.1.3 Size selection

Eluted Nebulised DNA was made up to 100µl with Buffer EB, and an appropriate, calibrated volume of AMPure Beads added. The mixture was vortexed and allowed to incubate at room temperature for five minutes, before being pelleted in a DynaMag 2 magnet and the supernatant removed. The beads were washed twice with 500µl 80% ethanol, and incubated for 30 seconds each time. The supernatant was removed again and the tubes were allowed to dry out in a heating block at 37°C. The dried pellets were resuspended in 24µl Qiagen's EB buffer and vortexed. They were pelleted by an MPC again, and the supernatant transferred to a fresh tube. The pellet was discarded.

9.1.1.4 DNA 7500 Bioanalyzer chip

An Agilent DNA 7500 Assay Kit and an Agilent 2100 Bioanalyzer were used for the determination of successful nebulisations and size selections. Manufacturer protocols were followed for the preparation of reagents, and loading of samples onto the chip (see section 2.2.5 p.78).

Successful samples showed a positive curve on the data with at least 80% of fragments having a length between 300-1000bp.

9.1.1.5 Fragment End Polishing

The following reagents were added to a PCR tube:

Nebulised DNA	~23µl
10x Polishing Buffer	5µl
BSA*	5µl
ATP*	5µl
dNTPs*	2µl
T4 PNK*	5µl
T4 DNA polymerase*	5µl
Total	50µl

* = Concentration of these reagents not provided by manufacturer

The tubes were capped securely, vortexed and spun down, then incubated on a PCR block at 12°C for 15 minutes, then at 25°C for a further 15 minutes. Polished fragments were purified using QIAquick PCR purification columns according to the manufacturer's protocol (see section 2.2.7.1 p.86), with the addition of a second dry spin after rotating the tubes 180°, and the elution of samples into 15µl of buffer EB.

9.1.1.6 Preparation of library immobilisation beads

50µl of Library Immobilisation Beads were transferred to a fresh microcentrifuge tube and pelleted using an MPC to remove the buffer. Beads were washed twice with 100µl of 2x Library Binding Buffer whilst still pelleted by the MPC. Beads were then resuspended in 25µl of 2 x Library Binding Buffer.

9.1.1.7 Adapter ligation

The following reagents were added to a PCR tube:

Polished DNA	~15µl
Ligase Buffer*	20µl
Adapters*	1µl
Ligase*	4µl
Total	40µl

* = Concentration of these reagents not provided by manufacturer

The tubes were capped securely, vortexed and spun down, then incubated on a PCR block at 25°C for 15 minutes. Polished fragments were purified using QIAquick PCR purification columns according to the manufacturer's protocol (see section 2.2.7.1 p.86), with the addition of a second dry spin after rotating the tubes 180°, and the elution of samples into 25µl of washed library immobilisation beads (see section 9.1.1.6 p.296).

9.1.1.8 Library immobilisation

Eluted adapter-ligated DNA and immobilisation beads were mixed well and placed on a tube rotator at room temperature for 20 minutes. The immobilised library was then washed twice with 100µl Library Wash Buffer using the MPC

9.1.1.9 Fill-In Reaction

The following reagents were added to the tube of washed immobilised library beads:

Molecular Biology Grade Water	40µl
10x Fill-in Polymerase Buffer	5µl
dNTP Mix*	2µl
Fill-in Polymerase*	3µl
Total	50µl

* = Concentration of these reagents not provided by manufacturer

The mixture was vortexed and incubated at 37°C for 20 minutes. The beads were then pelleted using an MPC and washed twice with 100µl of Library Wash Buffer.

9.1.1.10 Single-stranded template DNA Library Isolation

500µl of Qiagen's PBI buffer was mixed with 3.8µl of 20% acetic acid in a microcentrifuge tube to create a neutralisation solution. A melt solution was prepared by mixing 125µl of (10 N) sodium hydroxide with 9.875ml PCR grade water. 50µl of Melt solution was added to the Library-carrying beads and the tubes were vortexed. The supernatant was then transferred to the neutralisation solution using the MPC to pellet the beads. This process was repeated once. Extra acetic acid was added to the neutralisation solution as necessary to fully neutralise it and turn the solution yellow. The tube was then purified using QIAquick PCR purification columns according to the manufacturer's protocol (see section 2.2.7.1 p.86) and eluted into 15µl of TE buffer.

9.1.1.11 RNA Bioanalyzer

An Agilent RNA 6000 Pico Kit and an Agilent 2100 Bioanalyzer were used for the determination of successful library preps. Manufacturer protocols were followed for the preparation of reagents, and loading of samples onto the chip. Successful samples showed a positive curve

on the data, with a smaller adapter curve that was less than 5% the height of the library curve. Failed libraries showed either a larger adapter peak or no library peak at all.

9.1.2 Array capture

All reagents used for the array capture were supplied by NimbleGen. PCR reactions took place in a PCR room, separate from the pre-PCR room used for DNA library preparations.

9.1.2.1 Pre-LM PCR

A PCR Master Mix was prepared using the following volumes per sample:

5x GC-Rich PCR reaction buffer	50µl
25mM MgCl ₂	10µl
PCR Grade Nucleotide Mix	5µl
LM-PCR 454 Ti-A Oligo	5µl
LM-PCR 454 Ti-B Oligo	5µl
PCR-grade water	45µl
GC-Rich Enzyme Mix	5µl
Total	125µl

For each sample five PCR tubes were prepared containing ~3ng of DNA library in 25µl of PCR grade water and 25µl of Master Mix.

Pre-capture LM-PCR thermocycler program:

4 minutes at 94°C
11 cycles of:
30 seconds at 94°C
1 minute at 58°C
1.5 minutes at 68°C
End cycle
3 minutes at 68°C
Store at 4°C

The PCR tubes were removed from the thermocycler and the five tubes for each sample were combined into microcentrifuge tubes. QIAquick PCR purification columns were used to purify the samples according to the manufacturer's protocol (see section 2.2.7.1 p.86) with the following exceptions; 1250µl Qiagen Buffer PBI was added to each sample, and the buffer/sample mixed loaded in two halves between a spin. The washed DNA was eluted into 50µl PCR grade water and not buffer EB.

Eluted DNA was analysed to determine concentration and size distribution using a NanoDrop spectrophotometer and an Agilent 2100 Bioanalyzer with DNA 7500 chip and reagents according to the manufacturers' protocols (see sections 2.2.3.1 p.77 and 2.2.5 p.78 respectively). Successfully amplified samples had a PCR yield of at least 3.0µg, an A260/A280 reading between 1.7 and 2.0, and a Bioanalyzer trace that showed most fragments being between 350bp and 1000bp in length with an average fragment length of between 500-800bp.

9.1.2.2 NimbleGen Sequence Capture Array

The NimbleGen sequence capture array was used to capture target regions in our gene panel. The principle by which this method works is outlined in the introduction chapter. Capture hybridisation and array washing were both performed by Dr Morag Shanks.

9.1.2.3 Post-LM PCR

A PCR Master Mix was prepared using the following volumes per sample:

5x GC-Rich PCR reaction buffer	100µl
25mM MgCl ₂	20µl
PCR Grade Nucleotide Mix*	10µl
LM-PCR 454 Ti-A Oligo*	10µl
LM-PCR 454 Ti-B Oligo*	10µl
PCR-grade water	300µl
GC-Rich Enzyme Mix*	10µl
Total	460µl

* = Concentration of these reagents not provided by manufacturer

10 PCR reactions were prepared, each containing 4µl of captured DNA library.

9.1.3 Emulsion PCR

Emulsion PCR (emPCR, introduced in section 1.3.1 p.28) was performed in a PCR room, separate from the pre-PCR room and post-PCR areas. The emulsion break took place in a post-PCR room.

9.1.3.1 Preparation and Reaction

The emPCR reagents (excluding enzymes) were removed from cold storage one hour before use and allowed to thaw at room temperature. Once thawed the reagents were vortexed for five seconds. The tube of emPCR additive was vortexed and heated at 55°C for five minutes. All kit components including enzymes were then spun in a microcentrifuge for 10 seconds and the enzymes returned to ice.

The emulsion oil cups were secured to a TissueLyser and shaken at 28Hz for 2 minutes. 2ml of 5x Mock Amplification Mix was diluted with 8ml of Molecular Biology Grade Water and vortexed to make a working solution. The shaken emulsion oil cups were removed from the TissueLyser and 5ml of this working solution was added to each cup of Mock Amplification Mix. The cups were inverted two or three times and secured to the TissueLyser again. The cups were shaken at 28Hz for five minutes and removed from the TissueLyser.

A Live Amplification Mix was prepared using the following volumes:

- 2400µl Molecular Biology Grade Water
- 3000µl emPCR Additive*
- 1560µl 5x Amplification Mix
- 460µl Amplification primer*
- 400µl emPCR Enzyme Mix*
- 10µl PPIase*

* = Concentration of these reagents not provided by manufacturer

The Live Amplification Mix was vortexed and stored on ice.

1x Capture Bead Wash Buffer TW was prepared by mixing 1ml of 10x Capture Bead Wash Buffer TW with 9ml of molecular biology grade water. The tubes of DNA Capture beads were vortexed and the beads pelleted in a microcentrifuge by spinning for ten seconds, rotating 180° and spinning for a further ten seconds. The supernatant was discarded without disturbing the beads. Each tube of beads was washed twice with 1ml of 1x Capture Bead Buffer TW. The beads were vortexed to resuspend them, spun and the supernatant discarded.

An aliquot of the DNA library to be amplified was thawed. The library was added to the tubes of DNA Capture Beads using the following equation:

$$\mu\text{l of library per tube} = \frac{\text{desired molecules per bead} \times 35 \times 10^6 \text{ beads per tube}}{\text{Library concentration (molecules}/\mu\text{l)}}$$

The tubes were vortexed for five seconds and transferred to 15ml tubes. 3.75ml of the Live Amplification Mix was added to each tube. 1ml of the mix was pipetted to the DNA Capture Beads tubes that contained the captured library. These tubes were vortexed and transferred back to the 15ml tubes. This was repeated twice and the tubes vortexed.

Each 15ml tube was poured into a prepared emulsion cup. The emulsion cups were inverted three times and shaken in a TissueLyser at 12 Hz for five minutes.

The emulsified emPCR amplification mixes were then dispensed into 96 well PCR plates, with 100μl dispensed into each well. The wells were capped and properly sealed and placed on a thermocycler for the amplification reaction in the post-amplification area using the following program:

4 minutes at 94°C

50 x cycle:

30 seconds at 94°C

4.5 minutes at 58°C

30 seconds at 68°C

Final:

10°C on hold

9.1.3.2 Emulsion break

The Enhancing Fluid XT and Annealing Buffer XT were retrieved and stored at room temperature. A GS FLX Titanium emPCR Breaking Kit was assembled according to the manufacturer's instructions. A 50ml conical tube was attached to each of the two lids. The blue connector was attached to the 8-pronged transpette, and the other end of the tubing attached to a vacuum pump with liquid trap.

The emPCR plates were retrieved from the thermocycler and the assembled breaking kit was used to aspirate the emulsions from each emulsion cup, eight wells at a time, collecting them into the two conical tubes using slow circular motions of the transpette tips at the bottom of each well. After all the wells were aspirated the transpette was inverted to drain as much material as possible into the conical tubes.

The wells were rinsed twice with 100µl of isopropanol per well. The wells were aspirated in the same way each time, and the transpette again inverted at the end to aid drainage. After aspiration of the second rinse a further 5ml of isopropanol was aspirated slowly to collect any beads that remained in the tubing.

The vacuum pump was turned off and the two 50ml collection tubes carefully removed and re-capped with their original caps. Each tube was vortexed and the contents mixed back and forth between the tubes until the bead suspensions were of equivalent amounts of emulsion oil.

Isopropanol was added to each tube to make a final volume of 40ml. The tubes were vortexed. The beads were pelleted in a centrifuge at 930g for 5 mins. The supernatant was carefully poured off without disturbing the beads. This wash step was repeated with 35ml of Enhancing Fluid XT, then twice with 35ml of isopropanol, then 35ml of ethanol, and finally 35ml of Enhancing Fluid XT.

After the final wash step with Enhancing Fluid XT, 2ml was left in the tube. The DNA bead suspension was transferred to two 1.7ml tubes for each emulsion cap (four total) using a 1000µl pipette. The tubes were spun-rotated-spun and the supernatant discarded. This process was repeated until all the beads had been transferred to the four tubes. Each of the 50ml collection tubes was rinsed with 600µl of Enhancing Fluid XT, vortexed and this rinse added to the 1.7ml tubes, followed by a further spin-rotate-spin and discarding of the supernatant. 1ml of Enhancing Fluid XT was added to each tube and the pellets resuspended by vortexing.

A heating block was set to 65°C. A melt solution was prepared by mixing 125µl NaOH (10 N) with 9.875ml of Molecular Biology Grade Water. The tubes were spun-rotated-spun and the supernatant discarded.

1ml of Melt Solution was added to each tube. The tubes were incubated for two minutes at room temperature before again being spun-rotated-spun and the supernatant discarded. This melt step was repeated.

1ml of Annealing Buffer XT was added to each tube of beads and vortexed. The tubes were spun-rotated-spun and the supernatant discarded. This step was repeated.

45µl of Annealing Buffer XT and 25µl of Enrichment primer was added to each tube and vortexed. The tubes were placed in a heat block at 65°C for five minutes and then cooled on ice for two minutes.

800µl of Enhancing Fluid XT was added to each tube and vortexed. The tubes were spun-rotated-spun and the supernatant discarded. This step was repeated twice with 1ml of Enhancing Fluid XT. 800µl of Enhancing Fluid XT was added to each tube and vortexed.

The tube of brown Enrichment Beads was vortexed for one minute to resuspend its contents entirely. The Enrichment Beads were pelleted using a Magnetic Particle Collector (MPC). The supernatant was discarded carefully, so as not to disturb the beads.

1ml of Enhancing Fluid XT was added and vortexed. The Enrichment Beads were pelleted using an MPC and the supernatant carefully discarded so as not to disturb any of the beads. This step was repeated.

After discarding the supernatant, the tube was removed from the MPC. 320µl of Enhancing Fluid XT was added and vortexed.

80µl of washed Enrichment Beads were added to each tube of amplified DNA beads and vortexed. The tubes were then rotated on a Labquake at room temperature for five minutes. The tubes were placed in an MPC for 3 minutes to pellet the Enrichment Beads. The MPC was inverted several times. The supernatant was carefully discarded from each tube using a 1000µl pipette, taking care not to draw off any enrichment beads.

The following wash steps were then repeated until there were no visible beads remaining in the supernatant:

1. 1ml of Enhancing Fluid XT was added to each tube.
2. Tubes were removed from the MPC and vortexed.
3. The tubes were returned to the MPC to pellet the beads on the wall of the tubes.
4. The tubes were inverted.

5. Supernatant carefully discarded with a 1000µl pipette, taking care not to disturb the pelleted beads.

When white DNA beads were no longer being aspirated the tubes of enriched beads were removed from the MPC and resuspended in 700µl of Melt Solution. The tubes were vortexed for five seconds and the enrichment tubes placed in the MPC until the Enrichment Beads were pelleted. The supernatants from each enriched DNA beads tube were combined into a single 1.7ml collection tube. This tube was then spun-rotated-spun and the supernatant discarded.

700µl Melt solution was again added to the enrichment tubes. The tubes were vortexed for five seconds and the enrichment tubes placed in an MPC until the Enrichment Beads had pelleted. The supernatants containing the enriched DNA beads were transferred into the same 1.7ml collection tube. The enrichment tubes were discarded and the 1.7ml collection tubes spun-rotated-spun and the supernatants discarded.

1ml of Annealing Buffer XT was added to each collection tube and the supernatant discarded. This step was repeated twice. Each bead pellet was resuspended in 200µl of Annealing Buffer XT.

50µl of Sequencing Primer was added to each collection tube and vortexed. The tubes were placed in a heat block at 65°C for five minutes and then on ice for two minutes.

800µl of Annealing Buffer XT was added to each collection tube and vortexed for five seconds. The tubes were spun-rotated-spun and the supernatant discarded. This step was repeated two times with 1ml of Annealing Buffer XT.

1ml of Annealing Buffer XT was added to each bead pellet and vortexed. A 3µl aliquot of the beads was counted with a particle counter to calculate the percentage bead enrichment using the following equation:

$$\% \text{ Bead enrichment} = \frac{\text{Number of enriched beads}}{35 \times 10^6 \text{ beads/cup}} \times 100$$

The calculated % Bead enrichment should be between 5 and 20% for the library to yield good sequencing results. Beads were stored at 2°C to 8°C until sequencing.

9.1.4 Sequencing

The sequencing process was performed in a post-PCR area, separate from the pre-PCR room used for library preparations, and the PCR room used for amplifications.

The Sequencing Reagents Tray was removed from frozen storage. The Sequencing Enzymes sleeve was kept at -15°C to -25°C. The sequencing reagents tray was allowed to thaw by immersing in a water bath at room temperature for two hours. The Titanium Bead Buffer was kept on ice.

9.1.4.1 Pre-Wash

The previous sequencing run on the sequencer screen was closed if still left open. A pre-wash was performed by discarding the spent reagents and cleaning the reagents cassette. This involved opening the exterior fluidics door and raising the sipper manifold completely, sliding the Reagents cassette out, pouring out the fluid remaining in the container and discarding the spent bottles and tubes. The empty Reagent cassette was then thoroughly rinsed with warm tap water and all the outside surfaces dried with a paper towel. The used Sipper tubes were removed from the manifold and discarded. Gloves were then changed. New Sipper tubes were then inserted into the manifold and gloves were changed again.

The pre-wash cassette was prepared by placing the GS-FLX Pre-wash Tube Holder on top of the Reagents cassette and inserting 11 small tubes and four large tubes in the cassette. All the tubes were filled with Pre-wash buffer and the cassette was slid into the instrument. The manifold was lowered and the fluidics door closed.

The instrument tab was selected on the instrument computer. The start button was clicked in the global action area. The pre-wash option was selected on the Run Wizard's first window, and the start button selected in the second window.

9.1.4.2 Pico Titer Plate (PTP) Device Preparation

1.2ml of Titanium Supplement CB was added to the 200ml bottle of pre-chilled titanium bead buffer. The bottle was re-capped and mixed gently by inversion. The tube of apyrase (ATP-diphosphatase) was removed from the sequencing sleeve and completely thawed. The other components were left on ice. 34µl apyrase was added to the bottle containing Titanium bead buffer and Titanium supplement CB. The bottle was re-capped and labelled "bead buffer 2". The bottle was mixed gently by inversion and returned to ice, along with the remainder of the Titanium supplement CB and the apyrase.

The pico-titre plate (PTP) device (a plate containing micro fabricated wells into which sequencing beads are deposited) was removed from the shipping tray. The PTP device ID was noted down. The PTP device was completely submerged in BB2 for ten minutes. The bead loading gasket was washed thoroughly in MilliQ water and allowed to dry. The bead deposition device (BDD, a plastic device that holds the PTP in place while beads are loaded) was washed using MilliQ water and a bristle brush and allowed to dry.

9.1.4.3 Bead preparation

The Packing Beads were washed three times in 1ml of BB2 by centrifuging at 10,000rpm for five minutes. After the third wash 550µl of BB2 was added to each tube, the beads were resuspended and kept on ice.

The volume of DNA library beads needed was calculated based on the number of beads needed and the concentration of the library. The beads were vortexed and transferred to 2ml tubes. 20µl of Control DNA beads were added to each tube. The beads were spun for one minute at 10,000rpm. The tube was rotated and spun again for one minute. Supernatant was removed to leave a final volume of 50µl in each tube. The removed supernatant was discarded.

The DNA Bead Incubation Mix (DBIM) was prepared by mixing and vortexing the following:

BB2*	1570µl
Polymerase Cofactor*	150µl
DNA Polymerase*	300µl

* = Concentration of these reagents not provided by manufacturer

950µl of the DBIM mix was added to each tube and vortexed well. If the final volume of the tube was now less than 1000µl, additional DBIM was added to bring the total up to this volume. The tubes were then rotated on a lab rotator for fifteen minutes.

9.1.4.4 Enzyme and PPIase bead layers preparation

The Enzyme Beads and PPIase Beads were vortexed and pelleted using a Magnetic particle collector (MPC) for thirty seconds. The MPC was gently inverted several times and allowed to pellet again for thirty seconds. The supernatants were removed without disturbing the pellets and the tubes removed from the MPC. The Enzyme and PPIase Beads were washed three times with 1ml of BB2 for each tube of Enzyme Beads and with 500µl of BB2 for the tube of PPIase Beads using the MPC.

After the final wash 1ml of BB2 was added to each tube of Enzyme Beads and 500µl of BB2 to the tube of PPIase Beads. The tubes were vortexed and placed on ice.

Three tubes were prepared for layers 1, 3 and 4 of the Bead Deposition, made up with the following volumes:

Layer 1:	3250µl BB2*, 550µl Enzyme Beads*
Layer 3:	2500µl BB2*, 1300µl Enzyme Beads*
Layer 4:	3340µl BB2*, 460µl PPIase Beads*

* = Concentration of these reagents not provided by manufacturer

After the DNA bead incubation in DBIM was completed, 265µl Packing Beads and 435µl BB2 was transferred to the tube to make up bead layer two. The mixture was vortexed and rotated at room temperature for five minutes.

9.1.4.5 Bead Deposition

The Bead Deposition Device and PTP Device were assembled according to manufacturer's instructions. The beads were deposited onto the PTP device by injecting each layer through the port holes of the assembled BDD and then centrifuging the device for five minutes at 1620 x *g* to deposit the beads in the wells of PTP device. After each layer had been centrifuged the excess supernatant was gently removed with a pipette tip through a port hole of the BDD. During the centrifugation of each layer the next part of the protocol (removal of the pre-wash cassette) was performed.

9.1.4.6 Removal of the Pre-Wash Cassette

On the instrument computer a "sequencing run complete" message was displayed to signify the end of the pre-wash run. OK was clicked on the computer terminal of the machine. The fluidics door was then opened and the sipper manifold raised completely up. The Pre-wash tubes and Pre-Wash tube holder were removed from the Reagents cassette. The tubes were emptied in a sink and discarded. All the outside surfaces of the Reagents cassette were wiped with a paper towel. The fluidics area deck was wiped with 50% ethanol and a paper towel, taking care not to touch the sipper tubes. The deck was allowed to air dry completely. Gloves were changed and bead layer 2 was deposited at this point.

During the centrifugation of bead layer 2, 6.6ml of Titanium Supplement CB and 1000µl of DTT was added to each bottle of Titanium Buffer CB. The bottles were re-capped and gently inverted to mix. The sequencing reagents tray kept in the cold was placed in the right hand side of the Reagents cassette. 5µl of PPIase reagent was diluted into 45µl Inhibitor TW reagent in a 1.7ml tube and vortexed. A tablet of Sodium Chlorite was added to the "Post run wash tube" labelled with a white sticker. Gloves were changed. 13.2µl of diluted PPIase was added to the "Inhibitor

TW” labelled with a pink sticker. 260µl of Apyrase was added to the “Buffer for Apyrase” labelled with a yellow sticker. Gloves were changed. 3000µl of dATP was added to the “Buffer for dATP (A)” labelled with a lavender sticker. Gloves were changed. The lids were replaced on each tube and the tray inverted twenty times to mix the contents.

The caps were removed from all bottles and tubes and the Reagents cassette inserted into the instrument. The sipper manifold was carefully lowered and the exterior fluidics door closed. Gloves were changed and bead layer 3 was deposited at this point.

During the centrifugation of bead layer 3, the unlock button on the instrument tab of the computer terminal was clicked. The camera door was opened by pulling gently on the handle. The spent PTP device was removed from the cartridge by pressing the PTP frame spring latch and lifting the frame from the cartridge to slide out the used PTP device. The PTP frame was closed and a pair of plastic forceps used to remove the used PTP cartridge seal from the PTP cartridge.

A Kimwipe was wetted with 50% ethanol and used to wipe the surface of the cartridge to remove any bead and reagent residue. The cartridge was allowed to air dry completely. A Zeiss pre-moistened cleaning tissue was used to gently wipe the camera faceplate which was then also allowed to air dry completely. A Kimwipe was moistened with a 10% solution of Tween-20 and used to wipe the surface of the cartridge with.

9.1.4.7 Loading and Setting the Run Script and Other Run Parameters

The “Start” button was clicked on the global action area of the instrument computer. The “sequencing run” was selected on the next window. The IDs and barcodes noted down previously were entered and the “next” button pressed. A name for the run was entered and the “XLR 70 TI” option selected on the next window, followed by the “2 region” option on the window after that. On the next window “200 cycles” was selected. On the next window the data backup box was checked. On the next window no comments were made and the “next” button pressed. At this point the fourth bead layer was deposited.

After deposition of the fourth bead layer, the cartridge seal was placed on the cartridge of the machine with the grooved side facing up. The PTP device was removed from the BDD and the PTP slid into the cartridge frame facing down. The PTP frame was closed and the back of the PTP device wiped with a Kimwipe. The camera door was closed.

On the computer terminal of the machine the final screen now displayed a start button which was clicked to begin the sequencing run.

9.2 Cohort of patients with macular dystrophies

Table 9.1 Patients diagnosed with a Stargardt disease compatible phenotype in whom at least two potentially pathogenic variants in *ABCA4* have been identified.

Patient ID	Phenotype	Potential pathogenic variants identified		
		Exon/Intron	Nucleotide change	Protein change
183G1	STGD	Exon 5	c.551 C>T	p.S184F
		Exon 33	c.4793 C>A	p.A1598D
235J8	STGD	Exon 14	c.1957 C>T	p.R653C
		Exon 20	c.3004 C>T	p.R1002W
325E3	STGD	Exon 6	c.634 C>T	p.R212C
		Exon 30	c.4457 C>T	p.P1486L
390E1	STGD	Exon 44	c.6079 C>T	p.L2027F
		Exon 44	c.6079 C>T	p.L2027F
390F1	STGD	Exon 17	c.2588 G>C	p.G863A
		Exon 30	c.4469 G>A	p.C1490Y
502F5	STGD	Exon 29	c.4326 C>A	p.N1442K
		Exon 47	c.6449 G>A	p.C2150Y
503E6	STGD*	Exon 17	c.2588 G>C	p.G863A
		Intron 48	c.6730-1 G>C	
549H8	STGD	Intron 35	c.5018+2 T>C	
		Exon 46	c.6316 C>T	p.R2106C
554F5	STGD	Exon 30	c.4469 G>A	p.C1490Y
		Exon 37	c.5260 T>G	p.Y1754D
560E5	STGD	Exon 3	c.161 G>A	p.C54Y
		Exon 17	c.2588 G>C	p.G863A
604C1	STGD	Exon 19	c.2828 G>A	p.R943Q
		Exon 46	c.6316 C>T	p.R2106C
611A4	STGD	Exon 3	c.161 G>A	p.C54Y
		Exon 42	c.5882 G>A	p.G1961E
624F2	STGD	Exon 6	c.768 G>T	p.V256V
		Exon 22	c.3322 C>T	p.R1108C
658G5	STGD	Exon 5	c.550-551 TC>CG	p.S184R
		Exon 42	c.5882 G>A	p.G1961E
667H7	STGD	Exon 10	c.1253 T>C	p.F418S
		Exon 17	c.2588 G>C	p.G863A
673H3	STGD	Exon 42	c.5882 G>A	p.G1961E
		Intron 37	c.5313-3 C>G	

Table 9.1 Continued

Patient ID	Phenotype	Potential pathogenic variants identified		
		Exon/Intron	Nucleotide change	Protein change
680C6	STGD	Exon 20	c.2971 G>C	p.G991R
		Intron 28	c.4254-2 A>G	
681E3	STGD	Exon 28	c.4139 C>T	p.P1380L
		Intron 40	c.5714+5 G>A	
683C6	STGD	Exon 22	c.3322 C>T	p.R1108C
		Exon 44	c.6079 C>T	p.L2027F
676D5	STGD	Exon 30	c.4457 C>T	p.P1486L
		Exon 30	c.4457 C>T	p.P1486L
912D3	STGD	Exon 48	c.6709 A>C	p.T2237P
		Exon 21	c.3064 G>A	p.E1022K
503D6	STGD*	Exon 21	c.3056 C>T	p.T1019M
		Exon 22	c.3291 A>T	p.R1097S
		Exon 33	c.4753 C>T	p.R1585W
522A1	STGD	Exon 5	c.455 G>A	p.R152Q
		Exon 23	c.3386 G>T	p.R1129L
		Exon 30	c.4469 G>A	p.C1490Y
593F3	STGD	Exon 1	c.1 A>G	p.M1V
		Exon 17	c.2588 G>C	p.R943Q
		Exon 44	c.6089 G>A	p.R2030Q
775C5	STGD	Exon 5	c.455 G>A	p.R152Q
		Exon 28	c.4163 T>C	p.L1388P
		Exon 44	c.6079 C>T	p.L2027F
614C7	STGD	Exon 5	c.455 G>A	p.R152Q
		Exon 11	c.1411 G>A	p.E471L
		Exon 30	c.4537 dupC	p.Q1513fs
		Exon 46	c.6391 G>A	p.E2131K
729H3	STGD	Exon 13	c.1927 G>A	p.V643M
		Exon 18	c.2654 G>A	p.G885E
		Exon 20	c.2971 G>C	p.G991R
		Exon 24	c.3602 T>G	p.L1201R
		Exon 29	c.4322 G>A	p.G1441D
		Exon 49	c.6732 G>A	p.V2244V

* Patient not available for corroborative investigation

Table 9.2 Patients with macular dystrophy phenotypes other than STGD or FF.

Patient ID	Phenotype	Potential pathogenic variants identified		
		Exon/Intron	Nucleotide change	Protein change
245G1	FF	Exon 12	c.1715 G>A	p.R572Q
		Exon 17	c.2588 G>C	p.G863A
277F3	CRD (historical STGD)	Exon 31	c.4577 C>T	p.T1526M
		Intron 40	c.5714+5 G>A	
347G7	CRD (historical STGD)	Exon 38	c.5318 C>A	p.A1773E
		Exon 44	c.6079 C>T	p.L2027F
504E4	CRD (historical STGD)	Exon 28	c.4139 C>T	p.P1380L
		Exon 30	c.4462 T>C	p.C1488R
529G2	CRD (historical STGD)	Exon 12	c.1715 G>A	p.R572Q
		Exon 17	c.2588 G>C	p.G863A
586H5	CRD (historical STGD)	Exon 44	c.6118 C>T	p.R2040X
		Intron 40	c.5714+5 G>A	
724G1	BEM	Exon 6	c.768 G>T	p.V256V
		Exon 46	c.6319 C>T	p.R2107C
756C3	CRD (historical STGD)	Exon 28	c.4139 C>T	p.P1380L
		Exon 33	c.4670 A>G	p.Y1557C

FF = Fundus flavimaculatus, CRD = Cone-rod dystrophy, BEM = Bull's eye maculopathy.

10. Publications

CORRIGENDA

Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease

Morag E Shanks, Susan M Downes, Richard R Copley, Stefano Lise, John Broxholme, Karl AZ Hudspith, Alexandra Kwasniewska, Wayne IL Davies, Mark W Hankins, Emily R Packham, Penny Clouston, Anneke Seller, Andrew OM Wilkie, Jenny C Taylor, Jiannis Ragoussis and Andrea H Németh

European Journal of Human Genetics (2013) **21**, 1031; doi:10.1038/ejhg.2013.91

Correction to: *European Journal of Human Genetics* (2013) **21**, 274–280; doi:10.1038/ejhg.2012.172

The authors would like to apologise for the following errors in Table 2.

Three references cited in Table 2 were absent from the final reference list:

ABCA4 N965S

Sun H, Smallwood PM, Nathans J: Biochemical defects in ABCR protein variants associated with human retinopathies. *Nat Genet* 2000; **26**: 242–246.

CRX Y258X

Chau KY, Chen S, Zack DJ, Ono SJ: Functional domains of the cone-rod homeobox (CRX) transcription factor. *J Biol Chem* 2000; **275**: 37264–37270.

GUCY2D R838H

Wilkie SE, Newbold RJ, Deery E *et al*: Functional characterization of missense mutations at codon 838 in retinal guanylate cyclase correlates with disease severity in patients with autosomal dominant cone-rod dystrophy. *Hum Mol Genet* 2000; **9**: 3065–3073.

The dbSNP entry rs61749449 refers to N965D rather than N965S.

CRB1T, 745K should read CRB1, T745K.

Also, the following errors were found in Supplementary Table 2:

ABCA4 c.322C>T should read c.3322C>T

ABCA4 c.5882G>C should read c.5882G>A

ABCA4 c.6379C>T should read c.6391G>A

PRPF8 c.6926T>G should read c.6926A>C

ARTICLE

Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease

Morag E Shanks¹, Susan M Downes^{2,3}, Richard R Copley¹, Stefano Lise¹, John Broxholme⁴, Karl AZ Hudspith¹, Alexandra Kwasniewska^{4,5}, Wayne IL Davies^{3,6}, Mark W Hankins³, Emily R Packham⁷, Penny Clouston⁷, Anneke Seller⁷, Andrew OM Wilkie⁸, Jenny C Taylor¹, Jiannis Ragoussis⁴ and Andrea H Németh^{*,1,3,9}

Inherited retinal degeneration (IRD) is a common cause of visual impairment (prevalence ~1/3500). There is considerable phenotype and genotype heterogeneity, making a specific diagnosis very difficult without molecular testing. We investigated targeted capture combined with next-generation sequencing using Nimblegen 12plex arrays and the Roche 454 sequencing platform to explore its potential for clinical diagnostics in two common types of IRD, retinitis pigmentosa and cone-rod dystrophy. 50 patients (36 unknowns and 14 positive controls) were screened, and pathogenic mutations were identified in 25% of patients in the unknown, with 53% in the early-onset cases. All patients with new mutations detected had an age of onset <21 years and 44% had a family history. Thirty-one percent of mutations detected were novel. A *de novo* mutation in rhodopsin was identified in one early-onset case without a family history. Bioinformatic pipelines were developed to identify likely pathogenic mutations and stringent criteria were used for assignment of pathogenicity. Analysis of sequencing metrics revealed significant variability in capture efficiency and depth of coverage. We conclude that targeted capture and next-generation sequencing are likely to be very useful in a diagnostic setting, but patients with earlier onset of disease are more likely to benefit from using this strategy. The mutation-detection rate suggests that many patients are likely to have mutations in novel genes.

European Journal of Human Genetics (2013) **21**, 274–280; doi:10.1038/ejhg.2012.172; published online 12 September 2012

Keywords: retinal degeneration; molecular diagnostics; next-generation sequencing

INTRODUCTION

Inherited retinal degeneration (IRD) is a term used to describe a wide range of genetically determined retinal disorders, which cause varying degrees of progressive visual dysfunction and blindness. The nosology of IRD is complex, ranging from eponymous classifications to clinical and functional phenotyping using imaging, electrophysiology and psychophysical evaluation.

Many recognisable phenotypes are associated with mutations in specific genes, for example *ABCA4* in Stargardt disease, but retinitis pigmentosa and rod-cone/cone-rod dystrophies exhibit significant genetic heterogeneity (<https://sph.uth.tmc.edu/retnet> – last accessed 23 March 2012). These cases provide the greatest challenge for clinicians in providing accurate prognosis and genetic recurrence risks to family members. The frequency of mutations in each of these genes is unknown, and in a diagnostic setting sequential genetic testing might be considered. Hitherto, genetic testing technologies have been limited to Sanger sequencing, which is very accurate but prohibitively costly for large-scale sequencing, or screening methods

such as APEX arrays which detect only known mutations. Because of these complex clinical, technical and costing issues, very few genes have been available for testing within a clinical diagnostic setting in the UK or elsewhere (Supplementary Table 1).^{1,2} It is recognised that genetic testing has an increasingly important role in IRD, not only for clinical diagnostics and for prognosis but also for potential therapeutic interventions such as gene therapy, and since High Throughput or next-generation sequencing (NGS) allows multiple genes to be sequenced in parallel, this could be used to increase molecular diagnostic rates in IRD. As IRD is highly clinically and genetically heterogeneous, it offers an ideal model in which to investigate the complexities of introducing NGS into a clinical diagnostic setting. We therefore set out to identify pathogenic mutations in a cohort of patients with retinitis pigmentosa or cone-rod dystrophy from an Ophthalmic Genetic clinic to determine which patients would be most likely to benefit from a targeted screening approach in a clinical diagnostic setting. There have been two other recent publications using NGS in retinal dystrophies. The first

¹Oxford Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; ²Oxford Eye Hospital, John Radcliffe Hospital, Oxford University Hospitals NHS Trust, Oxford, UK; ³Nuffield Department of Clinical Neurosciences, Nuffield Laboratory of Ophthalmology, University of Oxford, Oxford, UK; ⁴Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; ⁵Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK; ⁶School of Animal Biology and University of Western Australia, Oceans Institute, University of Western Australia, Perth, Australia; ⁷Oxford NHS Regional Molecular Genetics Laboratory, Oxford University Hospitals NHS Trust, Oxford, UK; ⁸Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK; ⁹Department of Clinical Genetics, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK

*Correspondence: Dr AH Németh, Nuffield Department of Clinical Neurosciences, Nuffield Laboratory of Ophthalmology, University of Oxford, 6th Floor, West Wing, John Radcliffe Hospital, Oxford OX3 9DU, UK. Tel: +44 1865226020; Fax: +44 1865226011; E-mail: andrea.nemeth@eye.ox.ac.uk

Received 27 March 2012; revised 6 June 2012; accepted 5 July 2012; published online 12 September 2012

reported a detection rate of 57% in 20 patients with a mixture of retinal phenotypes and the second reported a 36% detection rate in 100 patients with retinitis pigmentosa. However, neither reported data on the age of onset of the condition in the patients who were analysed.^{3,4} We used targeted capture combined with NGS to enrich and sequence for genes of interest, to determine whether NGS could be a useful introduction into clinical practice and, if so, which patient groups would be most likely to benefit.

PATIENTS AND METHODS

Human subjects

All patients (and involved relatives) were counselled and consented according to the Declaration of Helsinki and the study was approved by the local Ethics Committee.⁵ Patients were recruited if they were diagnosed with retinitis pigmentosa (RP) or cone-rod dystrophy (CRD) phenotype in a specialist Ophthalmic Genetics Clinic (run by SMD). We excluded any patient from this analysis who did not have RP or CRD, or was considered to have a multifactorial condition such as age-related macular degeneration. We also excluded patients with likely Stargardt disease, who were analysed in a separate study.⁶

Fifty-five patients were targeted and sequenced, 50 of which were included in the final analysis (5 were excluded because of low coverage – see results). 14/50 samples were positive controls. The positive controls included 12 patients who had been screened for *ABCA4* mutations in a previous study and 2 patients with heterozygous dominant mutations identified in *PRPF8* and *PRPF31*.⁶ Nineteen pathogenic variants were known in the positive control group (Supplementary Table 2).

Targeted capture and next-generation sequencing

A custom NimbleGen 12-plex capture array was designed and manufactured by Roche NimbleGen. The array was designed to capture 1413 exons from 73 IRD genes and 1 deep intronic splice site in *CEP290*, known to be associated with LCA (c.2991 + 1655A>G) (Supplementary Table 3).⁷ The design also included the entire genomic region of *ABCA4* as some patients with a single mutation in *ABCA4* might harbour mutations in the intronic region, although only the 25 bp of intronic sequence surrounding the exons was included in the

current analysis. The target regions of interest (ROI), including exons and UTRs were identified and compiled using NCBI build 36.1, hg18 and the Ensembl database, and all annotated transcripts were included. The final NimbleGen design was 463.4 kb in size and was made up of 135 000 (135K) overlapping 60–90 bp probes that covered 96.4% ROI. Genes were included if they were known to have previously been associated with IRD in the literature or RetNet (<https://sph.uth.tmc.edu/retnet> – last accessed 23 March 2012).

Library preparations were made using the Roche manufacturer's protocol. In brief, 7 µg of patient's genomic DNA was sheared by nebulisation into fragments of 400–800 bp (mean 600 bp). The fragment ends were polished using T4 PNK and T4 DNA polymerase and then adaptors containing a unique 454 Multiple Identifier Sequence (MID) were ligated. The library was immobilised to streptavidin beads to allow elution of single-stranded DNA, which was amplified by linker-mediated PCR using adaptor-specific primers.

An amplified library of 1 µg was loaded onto 1 region of the 12-plex array with 100 µg of COT-1 and MID-specific hybridisation oligonucleotides. Each region of the array contained the full 135K probe design. The array was hybridised for 60–70 h, washed and eluted as per NimbleGen instructions. The captured DNA was then re-amplified using linker-mediated PCR.

Emulsion PCR was performed using an optimal ratio of DNA to capture beads as per manufacturer's instructions. The emulsion was broken, the capture beads containing DNA were recovered and the sequencing primers were added. These samples were sequenced on 2 × 454 Titanium runs: each run used the two-lane gasket option and the same 12-plex capture was run on both lanes.

Sequence-read mapping and variant calling

Read data were analysed using the 454 Sequencing System Software package (version 2.5.3). Sequencing run files (SFF format) were split by multiplex identifiers (MID) using the 'sfffile' program. The GS Reference Mapper software with default parameters was then used to map sequence reads to the human reference genome (NCBI build 36, hg18) and to detect variants. Identified variants were annotated through the refGene database and flagged if already present in dbSNP (build 130). Only high-confidence variants (ie, those reported in the '454HCDiff.txt' output file) within the targeted regions were considered.

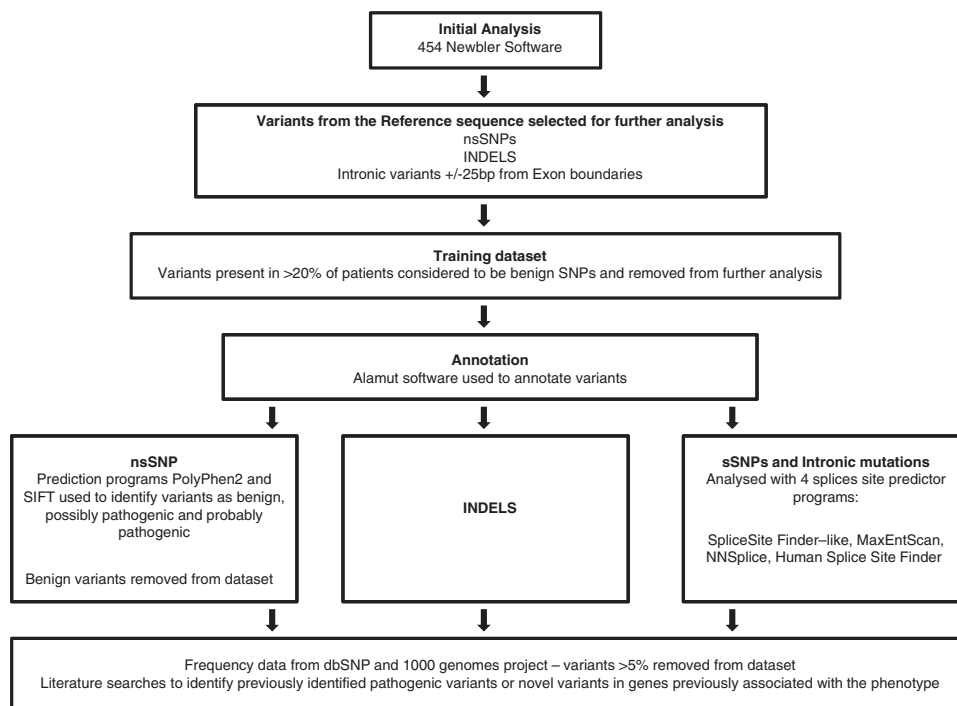


Figure 1 Data-analysis pipeline; the approximate number of variants at each analysis stage is shown.

Sanger sequencing validation

All mutations identified by NGS and considered likely to be pathogenic were validated by Sanger sequencing using standard protocols and where possible family members were tested to determine phase and investigate cosegregation of the variants with disease.

Sequence coverage was estimated from the '454ReadStatus.txt' output file with custom perl scripts. Only reads with mapping status 'Full' or 'Partial' were included in the analysis. The following metrics were determined: total data per array, percentage of total reads aligned to the ROI and for each ROI: minimum, maximum and mean number of reads and percentage of bases with $\geq 10\times$, $20\times$, $30\times$ coverage. These data were used to determine whether there was any bias of coverage in samples with or without mutations identified.

Variants were annotated using the analysis pipeline detailed in Figure 1. This pipeline included pathogenicity prediction programs PolyPhen2, SIFT and MutPred in addition to splice prediction programs MaxEntScan, NNSplice, Human Splice site finder and Splice Site finder.^{8–13} Missense mutations were considered likely to be pathogenic if (a) PolyPhen2, SIFT and MutPred all suggested pathogenicity (b) the variant has been previously reported with good evidence for a functional effect (c) low frequency in the general population and (d) in recessive cases the second mutation was found. We took a very conservative approach to considering missense mutations as pathogenic, because we have data from the analysis of rhodopsin to suggest that previously reported pathogenic mutations are in fact benign polymorphisms.¹⁴ Patient data were analysed by age of onset, presence or absence of family history and the phenotype. We calculated the percentage of bp of submitted exons that were missing from the NimbleGen final probe design. We also identified the number and percentage of exons missing from the final sequence data. We also analysed four samples with equivalent total reads and alignment to investigate the contribution of capture efficiency to the variability in overall depth of coverage obtained.

RESULTS

Samples of 55 patients were captured on a total of 8 arrays and sequenced in a total of 17 runs. In 5 of the 55 samples $<80\%$ of bases were covered at $10\times$ and the samples were excluded from further analysis. Data metrics for the remaining 50 samples (36 unknown IRD, 14 controls) are shown in Table 1.

Identification of mutations in positive controls

Eighteen of the 19 pathogenic mutations present in the 14 control samples were detected (Supplementary Table 2). The mutation that was not detected, *ABCA4* c.4537dupC in exon 30, is within a short homopolymer tract of 7 C (Supplementary Figure 1). The coverage of this base in this sample was $38\times$.¹⁵

Identification of variants considered likely to be pathogenic in patients with an unknown retinal degeneration

Mutations that were considered highly likely to be pathogenic, based on our analysis pipeline, were found in 9/36 patients in a total of 6 genes, making a total pathogenic mutation-detection rate of 25% (Table 2). In the patients with an age of onset under 21 the detection rate was 53% (Supplementary Table 5). The mutations were located at

13 different positions, of which 4 were novel. In total, 6 of the 9 patients had two autosomal recessive mutations, and 3 of the 9 patients had 1 autosomal dominant mutation. Surprisingly, 1 early-onset patient had the mutation P53R in Rhodopsin, which has only been described once before, segregating in an autosomal dominant RP family with adult-onset.¹⁶ This mutation was shown to be *de novo*, when parental samples were analysed (Table 2 and Supplementary Table 4) and analysis of photopigment function and cellular trafficking confirmed the pathogenicity of this variant.¹⁴ There is one very recent report of a *de novo* mutation in rhodopsin, although in that report, the data were inconclusive because paternal DNA was unavailable.⁴ Of the presumed pathogenic variants identified, 31% were present in dbSNP, demonstrating that this database likely contains numerous pathogenic variants. A further patient had a single mutation: 2465fsX2 in exon 54 of *CEP290* (which causes autosomal recessive disease), which was considered highly likely to be pathogenic, but the second mutation was not found. Overall, in the mutation-positive patients only 5 of the 13 mutations were missense mutations, reflecting the very conservative assessment we made of pathogenicity.

All patients with identifiable mutations had an age of onset <21 years and 44% had a positive family history: 2 with a dominant family history, and 2 with a family history suggestive of recessive inheritance (affected siblings or consanguinity). In the patients without identifiable mutations, 70% had an age of onset over 21 years and 78% did not have any relevant family history (Figure 2 and Supplementary Table 5).

In the patient with the *CEP290* deletion in exon 54 that produced a premature truncation codon (Table 2), no second mutation could be found, even in the cryptic splice site. The coverage for *CEP290* in this patient was over 99% of bases covered by at least $10\times$ and 82% covered by at least $30\times$, suggesting that low coverage is not the explanation for the missing second mutation.

When the records of all the cases were reviewed retrospectively, the mutations were considered to be consistent with the phenotype. A total of 19 variants were checked by Sanger sequencing of which three were not verified. Two had a low ratio of variant to reference sequence, and the third was an insertion, suggesting that all were artefacts.

Capture and sequence coverage variability

In all, 4.6% of exons, including 13 full exons, were not covered by the NimbleGen probes. However, we found that not all the regions were consistently missing from the sequence data, presumably because some of the regions are captured by neighbouring probes. One exon in *PITPNM3* was consistently and entirely missed from the final data despite being present in the design. We compared 4 samples with similar levels of coverage ($61\text{--}71\times$) and found that there was considerable variation in the base pairs and exons that were missing (Table 3). For example: 93–97% of base pairs were covered at $10\times$ and 94–98% of exons were completely covered ($10\times$ or more). However, if coverage at $30\times$ is considered as the gold standard, then

Table 1 Overall sequencing metrics.

Data (Mb)		Alignment (%)		% bp coverage $>10\times$		Coverage (%)		% bp coverage $>30\times$		Base coverage per sample	
Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
430	322–570	58	10–76	94	82–99	85	62–97	71	33–97	58	24–117

Abbreviation: bp, base pair.

Table 2 Presumed pathogenic variants

Patient	Diagnosis	Age at onset	Variant AA(s)	dbSNP	Evidence for pathogenicity for novel mutations	Family history	Co-segregation	Reference
IRD_111	RP	3	RHO, P53R	rs28933395	P53R expressed at low levels, regenerated chromophore poorly and was retained in the ER	Simplex	Mutation <i>de novo</i>	14,16
IRD_092	Cone/rod dystrophy	8	ABCA4, Q2220X ABCA4, Q2220X	—	A PTC mutation that results in the loss of 53AA	2 affected male cousins. Consanguineous first cousin parents	Parents heterozygous for mutation	17
IRD_086	RP	10	ABCA4, N965S	rs61749449	Sun <i>et al</i> showed that N965S mutation results in a protein that shows very little retinal-stimulated ATP hydrolysis	Affected sister consanguineous first cousin parents	Sister positive for homozygous mutation	18
IRD_082	RP	11	ABCA4, E636X	—	PTC that results in the loss of almost 70% of the ABCA4 protien.	Simplex	—	—
IRD_123	RP	14	ABCA4, V256V PDE6B, E55X ^a	—	Results in a predicted splice-site mutation PTC that results in the loss of 90% of AA including the cGMP binding domain and the catalytic domain	Simplex	—	19,20 21
IRD_094	Cone/rod dystrophy	15	PDE6B: I18 c.2193+1 GUCY2D, R838H	— rs61750173	Results in a predicted splice-site mutation Mutant has been shown to be less sensitive to calcium concentrations and is catalytically active even at high calcium levels. This activity probably causes excess calcium to build up in photoreceptor eventually causing apoptosis.	Affected daughter	—	22
IRD_121	Macular dystrophy	17	CRX, 258X ^a	—	PTC results in the loss of the OTX domain at the carboxy terminus (AA 284–295). A study by Chau <i>et al</i> showed this loss results in a mutant protein that, although still able to bind to CRX recognition sites, was unable to initiate the rhodopsin promoter	3 affected siblings	All 3 siblings positive for mutation-	—
IRD_095	RP	18	PDE6B: Q298X PDE6B: I18, c.2194-1 ^a	—	PTC causing the loss of 35% of the protein Predicted by software to result in a splice-site mutation	Simplex	—	23
IRD_100	LCA;severe RP	4 months	^a CRB1T, 745K CRB1, c.2842+5	rs28939720 (dbSNP entry for T745M) —rs28939720	T745M previously reported and found to disrupt an area of cytoplasmic domain that was found to be crucial to function in drosophila Predicted by software to result in a splicesite mutation	Simplex	—	24,25

Abbreviations: AA, amino acid; dbSNP, single-nucleotide polymorphism database; IRD, inherited retinal degeneration; PTC, premature truncation codon.

^aNovel mutation not previously described.

only 71–82% of exons were covered at this depth. Although there was a general correlation between missing exons and depth of coverage, this was not linear: our conclusions from this analysis are that for a specific depth of coverage, there is significant variability in the ROI with missing data.

Analysis comparing different patient-group metrics

When the data were analysed for percentage coverage at <10, <20 and <30 ×, we found that there was no evidence of bias towards the identification of pathogenic mutations in regions of greater coverage (Figure 3). The positive control samples showed a similar spread of coverage compared with the probands, suggesting that there is no specific bias in our data set to account for some patients not having mutations detected (Figure 3).

DISCUSSION

We used targeted NGS in patients with retinitis pigmentosa and cone-rod dystrophy to address the issues specifically related to clinical diagnostics, namely: (a) the characteristics of patients most likely to have mutations detected in order to provide guidelines for genetic testing, (b) genotype–phenotype correlations, (c) determining important sequencing metrics, and (d) bioinformatic challenges of determining pathogenicity in the context of large number of sequence variants. These issues are particularly relevant as targeted sequencing is likely to be the technology of choice for diagnostic purposes. This is because whole exome/genome-analysis pipelines need to be refined further to provide the accuracy required in a diagnostic setting.

We found that certain patient categories increased the likelihood of detecting mutations in our gene set. The most important was the age

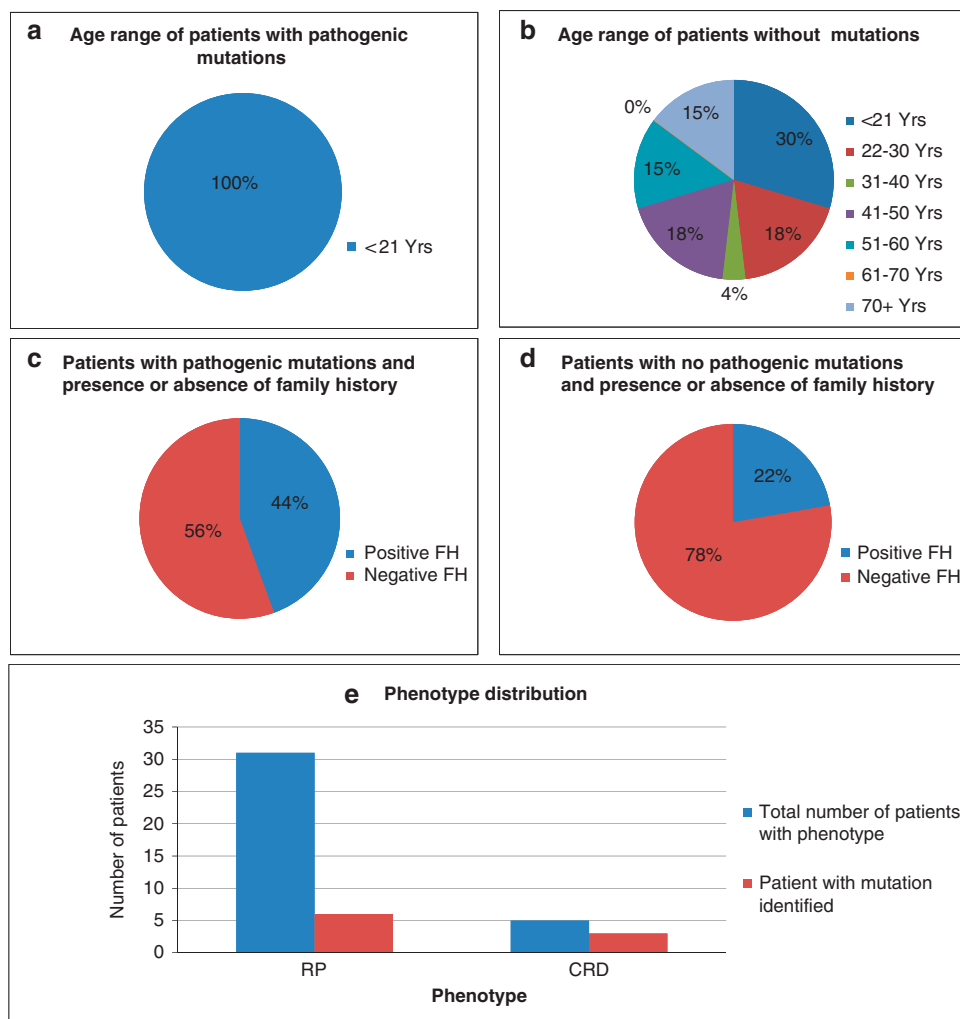


Figure 2 Age range and family history status of patients. (a) Age range of patients with pathogenic mutation(s) identified. (b) Age range of patients without pathogenic mutation(s) identified. (c) Family history in patients with pathogenic mutations identified (d) Family history in patients without a pathogenic mutation identified (e) Distribution of phenotypes in patients with a pathogenic mutation (red) compared with the total (blue).

Table 3 Capture variability. Metrics for four samples normalised for number of reads, average coverage and alignment

Sample	Alignment to ROI (%)	% bp covered at > 10×	% bp covered at > 30×	Av coverage	No of mapped reads	No of ROI at 10×	% of ROI at 10×	No of ROI with 20% miss at 10×	% of ROI with 20% miss at 10×	No of ROI at 30×	% of ROI at 30×
IRD_066	71	93	74	66	150817	1336	94	1316	93	1004	71
IRD_059	70	94	76	64	142196	1353	96	1342	95	1014	72
IRD_077	55	96	79	61	143850	1375	97	1372	97	1103	78
IRD_129	71	97	82	71	152325	1390	98	1387	98	1164	82

Abbreviations: bp, base pair; IRD, inherited retinal degeneration; ROI, regions of interest. Column 7 indicates all ROI in original array design.

of onset: all the patients with mutations had an onset under 21 years of age. This skewing of mutation detection towards the earlier age of onset group in our patients suggests that routine screening of patients over the age of 50 in a diagnostic setting is probably not warranted until further data are available. The explanation for skewing towards cases with an onset of under 21 years of age is unclear, but it may be that cases with an onset of over 50 years of age have a multigenic or multifactorial aetiology including accumulated environmental factors.

A positive family history was a poorer predictor in identifying a mutation. Although simplex cases are often considered most likely to represent autosomal recessive disease, we found a *de novo* dominant mutation in one patient, illustrating that it is important to include this possibility when providing genetic counselling. 8/17 (47%, Supplementary Table 5) cases with an onset of under 21 *did not* have mutations detected in contrast to the 53% of who did. Further work is in progress to investigate this, using whole-exome sequencing to determine if these patients have mutations in novel genes.

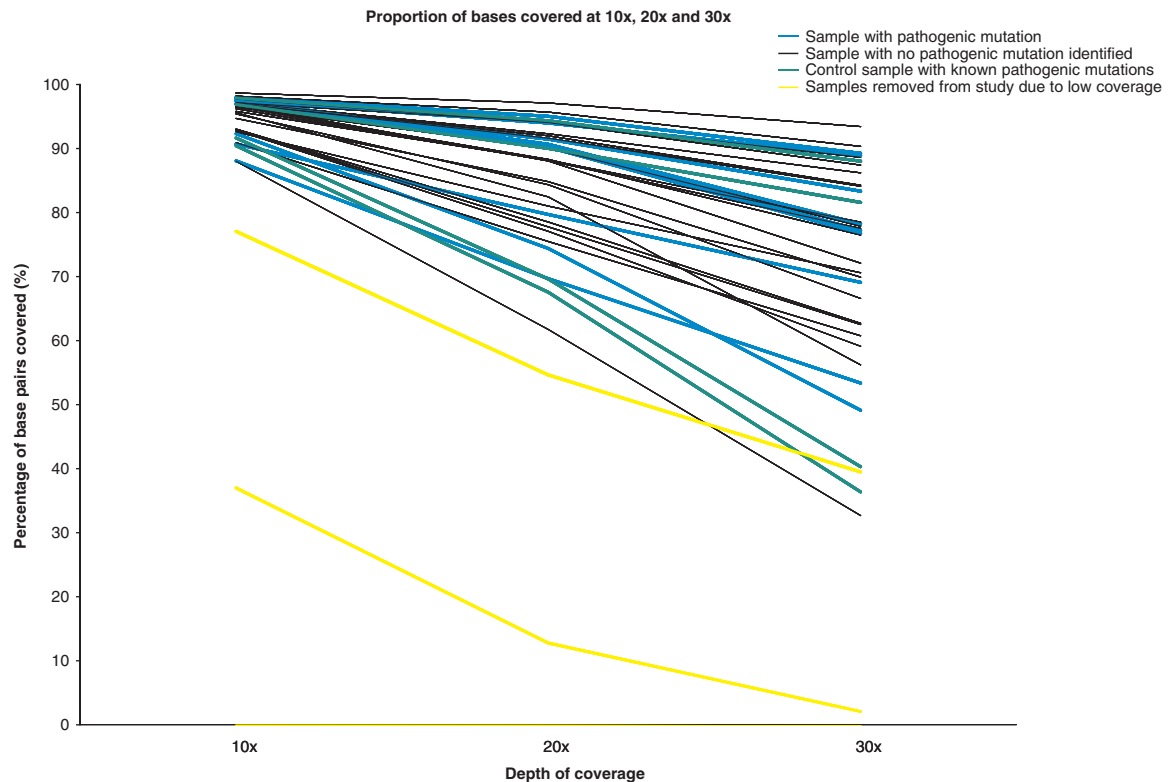


Figure 3 Sample coverage. The percentage of bases covered at $10\times$, $20\times$ and $30\times$ for each patient. Patients with mutation(s) identified (blue), patients with no identified mutations (black), control patients (green) and patients with $<80\%$ of bases covered at $10\times$ (yellow).

Although the phenotype was retrospectively considered to be consistent with the mutations found, this could not have been predicted *a priori*, illustrating the benefits of sequencing multiple genes in parallel. This was particularly the case for *ABCA4* mutations in which the patients had atypical non-STGD phenotypes, a well-recognised phenomenon with *ABCA4* retinopathies.¹⁷ We found 1 patient with a single-*CEP290* mutation. Mutations in this exon are known to cause Joubert and Meckel syndromes and LCA, so we considered this mutation highly likely to be pathogenic, but we were unable to identify the 2nd mutation despite good coverage across the gene.^{7,26} Thirty-one percent of the mutations we identified were novel, indicating that screening methods, which only detect for known mutations, would have missed a significant number of these in our data set.

We examined metrics that will be relevant when introducing NGS into clinical diagnostics. In particular, we examined metrics that might affect accuracy of the result, given that in a diagnostic setting this is of crucial importance. We found that there was variability in the capture efficiency, reflected by the variability of alignment ($10\text{--}76\%$, average 58%), which is similar to other reports of hybridisation-based capture.²⁷ We also found that average depth of coverage was highly variable ($24\text{--}117\times$, average $58\times$). Even when we controlled for the coverage, we found that the ROI that were missing were highly variable and not consistent, with only one exon that was entirely missed in all samples. A specific problem we noted was the difficulty of detecting mutations in homopolymer tracts using the 454 pyrosequencing chemistry. We show here that this is a problem with the 454 software rather than the chemistry itself, as the mutation was detected, but failed on quality scores. However, altering the software to increase mutation detection will increase the number of false positives leading to the need for additional validation by Sanger sequencing.

A major consideration for diagnostics is determining pathogenicity in the context of large number of sequence variants. Our analysis pipeline required several modifications during the analysis process. We found that if we excluded variants present in dbSNP during the filtering process, we removed 31% of the likely pathogenic variants in our data set. This applied to both recessive and dominant mutations. We also found that when synonymous variants were excluded, a splice mutation was missed. Therefore we adjusted our filtering algorithms to include splice predictions for synonymous variants and to include variants in dbSNP. We also noted that some indels were miscalled by the 454 software as synonymous SNPs and these were only identified manually. Use of commercially available software programmes may resolve some of these issues.

To assess pathogenicity of individual variants, we used standard pathogenicity prediction programmes, and missense variants that were considered to be benign by any of the programs were removed. This strategy reduced the number of variants to be analysed, but may have resulted in some pathogenic variants being removed and therefore missed during the filtering process. We were particularly conservative in our assessment of missense mutations, as we have recently found missense mutations in rhodopsin that are classified and reported as pathogenic by prediction programs but are almost certainly benign polymorphisms.¹⁴ However, this is not a new problem, as it is well recognised that bioinformatics prediction programs have limitations.

Two other papers have published data on the use of NGS for diagnostics in retinal degeneration with detection rates varying from 25% in our current study to 57% . These differences in detection rates are almost certainly due to variations in patient selection criteria, although it is possible that our lower detection rate is also due to our very conservative approach in making estimates of pathogenicity.

Specifically, Neveling *et al*,⁸ considered missense variants to be pathogenic, if two of the three programs predicted pathogenicity, whereas we required all three programs to consider a missense variant to be pathogenic. These differences emphasise that patient-selection criteria are of utmost importance in maximising the detection rate in a diagnostic setting and also that interpretation of pathogenicity remains a challenge for diagnostic services.¹⁴

CONCLUSIONS

We have shown that NGS offers significant advantages over conventional Sanger sequencing or screening technologies for clinical diagnostics of RP/CRD. However, we also show that capture methods, sequencing chemistries and bioinformatics pipelines need to be optimised to use NGS in a clinical diagnostic setting with confidence. Until these are achieved, the following points should be considered:

- Genetic testing guidelines and strategies to determine which patients are most likely to benefit from testing will become clearer as NGS use becomes more widespread. Our data suggest that targeting testing towards those with a younger age of onset or a family history may be the most cost effective. When reporting that 'no mutation has been identified', it should be made clear that some mutations may not be detected owing to factors such as lack of depth of coverage and miscalling of bases; as NGS evolves these problems will be addressed.
- All variants should be validated by Sanger sequencing before they are formally reported in the clinical setting.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Oxford Partnership Comprehensive Biomedical Research Centre with funding from the Department of Health's NIHR Biomedical Research Centre Programme. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health. This work was also supported in part by Roche Diagnostics, AK was funded by Ataxia UK, and JB and JR were funded by The Wellcome Trust (grant 075491/Z/04). WILD was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) grant to MWH.

- 1 Jaakson K, Zernant J, Külm M *et al*: Genotyping microarray (gene chip) for the ABCR (ABCA4) gene. *Hum Mutat* 2003; **22**: 395–403.
- 2 Klevering BJ, Yzer S, Rohrschneider K *et al*: Microarray-based mutation analysis of the ABCA4 (ABCR) gene in autosomal recessive cone-rod dystrophy and retinitis pigmentosa. *Eur J Hum Genet* 2004; **12**: 1024–1032.
- 3 Audo I, Bujakowska K, Leveillard T *et al*: Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases. *Orphanet J Rare Dis* 2012; **7**: 8.

- 4 Neveling K, Collin RW, Gilissen C *et al*: Next generation genetic testing for retinitis pigmentosa. *Hum Mutat* 2012; **33**: 963–967.
- 5 World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *J Postgrad Med* 2002; **48**: 206–208.
- 6 Downes SM, Packham ER, Cranston T, Clouston P, Seller A, Nemeth AH: Detection rate of pathogenic mutations in ABCA4 using direct sequencing: clinical and research implications. *Arch Ophthalmol* (in press).
- 7 den Hollander AI, Koenekoop RK, Yzer S *et al*: Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet* 2006; **79**: 556–561.
- 8 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
- 9 Ng PC, Henikoff S: Predicting deleterious amino acid substitutions. *Genome Res* 2001; **11**: 863–874.
- 10 Li B, Krishnan V, Mort ME *et al*: Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009; **25**: 2744–2750.
- 11 Yeo G, Burge CB: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004; **11**: 377–394.
- 12 Reese MG, Eeckman FH, Kulp D, Haussler D: Improved splice site detection in Genie. *J Comput Biol* 1997; **4**: 311–323.
- 13 Desmet FO, Hamroun D, Lalande M, Colod-Bérout G, Claustres M, Bérout C: Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009; **37**: e67.
- 14 Davies WIL, Downes SM, Fu JK *et al*: Next Generation Sequencing (NGS) in healthcare delivery: lessons from the functional analysis of rhodopsin. *Genet Med* 2012; e-pub ahead of print 12 July 2012; doi:10.1038/gim.2012.73.
- 15 Margulies M, Egholm M, Altman WE *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**: 376–380.
- 16 Inglehearn CF, Keen TJ, Bashir R *et al*: A completed screen for mutations of the rhodopsin gene in a panel of patients with autosomal dominant retinitis pigmentosa. *Hum Mol Genet* 1992; **1**: 41–45.
- 17 Maugeri A, Klevering BJ, Rohrschneider K *et al*: Mutations in the ABCA4 (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *Am J Hum Genet* 2000; **67**: 960–966.
- 18 Shroyer NF, Lewis RA, Lupski JR: Complex inheritance of ABCR mutations in Stargardt disease: linkage disequilibrium, complex alleles, and pseudodominance. *Hum Genet* 2000; **106**: 244–248.
- 19 September AV, Vorseter AA, Ramesar RS, Greenberg LJ: Mutation spectrum and founder chromosomes for the ABCA4 gene in South African patients with Stargardt disease. *Invest Ophthalmol Vis Sci* 2004; **45**: 1705–1711.
- 20 Zernant J, Schubert C, Im KM *et al*: Analysis of the ABCA4 gene by next-generation sequencing. *Invest Ophthalmol Vis Sci* 2011; **52**: 8479–8487.
- 21 McLaughlin ME, Ehrhart TL, Berson EL, Dryja TP: Mutation spectrum of the gene encoding the beta subunit of rod phosphodiesterase among patients with autosomal recessive retinitis pigmentosa. *Proc Natl Acad Sci USA* 1995; **92**: 3249–3253.
- 22 Weigell-Weber M, Fokstuen S, Török B, Niemeyer G, Schinzel A, Hergersberg M: Codons 837 and 838 in the retinal guanylate cyclase gene on chromosome 17p: hot spots for mutations in autosomal dominant cone-rod dystrophy? *Arch Ophthalmol* 2000; **118**: 300.
- 23 McLaughlin ME, Sandberg MA, Berson EL, Dryja TP: Recessive mutations in the gene encoding the α -subunit of rod phosphodiesterase in patients with retinitis pigmentosa. *Nat Genet* 1993; **4**: 130–134.
- 24 den Hollander AI, ten Brink JB, de Kok YJ *et al*: Mutations in a human homologue of *Drosophila crumbs* cause retinitis pigmentosa (RP12). *Nat Genet* 1999; **23**: 217–221.
- 25 den Hollander AI, Davies J, van der Velde-Visser SD *et al*: CRB1 mutation spectrum in inherited retinal dystrophies. *Hum Mutat* 2004; **24**: 355–369.
- 26 Brancati F, Barrano G, Silhavy JL *et al*: CEP290 mutations are frequently identified in the oculo-renal form of Joubert syndrome-related disorders. *Am J Hum Genet*, **81**: 104–113.
- 27 Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ: Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics. *Clin Chem* 2010; **56**: 1297–1306.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)

Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis

Hilary C. Martin^{1,†}, Grace E. Kim^{2,†}, Alistair T. Pagnamenta^{1,3}, Yoshiko Murakami⁴, Gemma L. Carvill⁵, Esther Meyer^{6,7}, Richard R. Copley^{1,3}, Andrew Rimmer¹, Giulia Barcia⁸, Matthew R. Fleming², Jack Kronengold², Maile R. Brown², Karl A. Hudspeth^{1,3}, John Broxholme¹, Alexander Kanapin¹, Jean-Baptiste Cazier¹, Taroh Kinoshita⁴, Rima Nabbout⁸, The WGS500 Consortium[‡], David Bentley⁹, Gil McVean¹, Sinéad Heavin¹⁰, Zenobia Zaiwalla¹¹, Tony McShane¹², Heather C. Mefford⁵, Deborah Shears¹³, Helen Stewart¹³, Manju A. Kurian⁶, Ingrid E. Scheffer¹⁰, Edward Blair¹³, Peter Donnelly^{1,¶}, Leonard K. Kaczmarek^{2,¶} and Jenny C. Taylor^{1,3,¶,*}

¹Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK, ²Departments of Cellular and Molecular Physiology and Pharmacology, Yale University School of Medicine, New Haven, CT, USA, ³NIHR Biomedical Research Centre, Oxford, UK, ⁴Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Osaka, Japan, ⁵Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, WA, USA, ⁶Neurosciences Unit, UCL-Institute of Child Health, London, UK, ⁷Department of Neurology, Great Ormond Street Hospital, London, UK, ⁸Department of Paediatric Neurology, Centre de Référence Epilepsies Rares, Hôpital Necker-Enfants Malades, Paris, France, ⁹Illumina Inc., San Diego, CA, USA, ¹⁰Departments of Medicine and Paediatrics, Florey Institute, The University of Melbourne, Austin Health and Royal Children's Hospital, Melbourne, VIC, Australia, ¹¹Department of Clinical Neurophysiology, John Radcliffe Hospital, Oxford, UK, ¹²Department of Paediatrics, Children's Hospital Oxford, John Radcliffe Hospital, Oxford, UK and ¹³Department of Clinical Genetics, Oxford University Hospitals NHS Trust, Oxford, UK

Received September 20, 2013; Revised and Accepted January 20, 2014

In severe early-onset epilepsy, precise clinical and molecular genetic diagnosis is complex, as many metabolic and electro-physiological processes have been implicated in disease causation. The clinical phenotypes share many features such as complex seizure types and developmental delay. Molecular diagnosis has historically been confined to sequential testing of candidate genes known to be associated with specific sub-phenotypes, but the diagnostic yield of this approach can be low. We conducted whole-genome sequencing (WGS) on six patients with severe early-onset epilepsy who had previously been refractory to molecular diagnosis, and their parents. Four of these patients had a clinical diagnosis of Ohtahara Syndrome (OS) and two patients had severe non-syndromic early-onset epilepsy (NSEOE). In two OS cases, we found *de novo* non-synonymous mutations in the genes *KCNQ2* and *SCN2A*. In a third OS case, WGS revealed paternal isodisomy for chromosome 9, leading to identification of the causal homozygous missense variant in *KCNT1*, which produced a substantial increase in potassium channel current. The fourth OS patient had a recessive mutation in *PIGQ* that led to exon skipping and defective glycoposphatidyl inositol biosynthesis. The two patients with NSEOE had likely pathogenic *de novo* mutations in *CBL* and *CSNK1G1*, respectively. Mutations in these

*To whom correspondence should be addressed at: Genomic Medicine Theme, Oxford Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK. Tel: +44 1865287633; Fax: +44 1865287770; Email: jenny@well.ox.ac.uk

[†]These authors contributed equally to this work.

[‡]A full list of members is provided in the Supplementary Material.

[¶]These authors jointly directed this work.

genes were not found among 500 additional individuals with epilepsy. This work reveals two novel genes for OS, *KCNT1* and *PIGQ*. It also uncovers unexpected genetic mechanisms and emphasizes the power of WGS as a clinical tool for making molecular diagnoses, particularly for highly heterogeneous disorders.

INTRODUCTION

Many recent studies have successfully used whole-exome or whole-genome sequencing (WES, WGS) to uncover the genetic basis of rare disorders (reviewed by 1,2), primarily in a research context. In addition, WES and WGS offer potentially revolutionary approaches to molecular diagnosis for patients in a clinical setting. In order to assess the possible clinical utility of WGS, we have sequenced the genomes of 500 individuals with a variety of medical conditions, including cancer, immunological disease and rare, putatively monogenic syndromes (3–5). As part of this ‘WGS500 project’, we sequenced six patients with severe early-onset epilepsy who had been previously refractory to molecular diagnosis.

Severe early-onset epilepsy is a good candidate for WGS as it is a challenging disorder to understand mechanistically. It represents a broad spectrum of phenotypes which are highly heterogeneous at the clinical and molecular levels (6). While some causative genes have been identified for many of these sub-phenotypes, the limitations of current technologies mean that genetic testing is largely confined to the genes associated with the specific presenting phenotype. However, it is increasingly being recognized that a given gene can cause multiple phenotypes (6), and that more comprehensive genetic testing may improve molecular diagnostic yield (7). This is useful clinically not only because it can help make or confirm a diagnosis, but also because it may allow counseling on recurrence risk and prenatal testing.

In its most severe form, early-onset epilepsy involves frequent seizures beginning in the first three months of life, with abundant epileptic activity that contributes to significant cognitive and motor delay (6,8). It is frequently associated with gross structural brain abnormalities and occasionally with metabolic disorders, which are often genetic in origin (9). Ohtahara Syndrome (OS) is a severe form of early-onset epilepsy characterized by a distinctive electroencephalogram (EEG) pattern known as ‘burst-suppression’, which consists of periodic high voltage bursts of slow waves mixed with spikes, followed by marked attenuation (10). The frequency of OS is about 1 in 100 000 live births (11). Children with OS typically have multiple seizure types including tonic spasms and focal seizures, which are often refractory to anti-epileptic drugs (12). Affected children may progress onto other epilepsy syndromes such as West Syndrome (6), or they may die in infancy.

Several genes have been implicated in severe early-onset epilepsy. The first reported for OS was the X-linked *ARX* gene, which encodes a developmental transcription factor (13). *De novo* mutations in *STXBPI* (14,15), encoding a protein involved in synaptic vesicle release, in *CDKL5* (16), encoding a serine/threonine kinase, and in ion channel genes *KCNQ2* (17), *SCN2A* (18,19) and *SCN8A* (20) have also been implicated, as have recessive mutations in the glutamate transporter *SLC25A22* (21). However, many OS patients test negative for

mutations in these genes, indicating that other genes have yet to be identified. Multiple additional genes have been associated with the broader range of early-onset epilepsies, including genes encoding cytoskeletal components [*SPTAN1* (22)] and proteins involved in signaling [*PLCB1* (23)], DNA repair [*PNKP* (24)] and neurotransmitter synthesis [*PNPO* (25)]. However, clinical testing is limited by the availability and costs of conventional single-gene tests, and thus tends to be restricted to genes that have been associated with the specific type of epilepsy. There is therefore scope to apply a more comprehensive approach to diagnosis using whole-genome methods.

In this study, we sequenced six patients with sporadic severe early-onset epilepsy, and their healthy parents. The patients were selected because traditional clinical molecular genetic approaches had failed to uncover the causal mutation. Four of the children had been diagnosed with OS, and two had severe non-syndromic early-onset epilepsy (NSEOE). Because these six cases were all sporadic, and the families were reported as non-consanguineous, we anticipated that the causal mutation was most likely to be *de novo*, but we also considered the simple, compound and X-linked recessive models.

RESULTS

We sequenced the six trios (Table 1; Supplementary Materials, Note S1) to high coverage on the Illumina HiSeq platform. In searching for the causal mutations, we considered coding variants as well as variants in regulatory regions within 50 kb of known early-onset epilepsy genes (see Materials and Methods). The most plausible causal variant in each trio was a coding mutation, and we report these here. See the Supplementary Materials, Note S2 for a description of candidate variants that were not deemed to be causal.

Patients 1 and 3: *KCNQ2* and *SCN2A*

Two OS cases had *de novo* non-synonymous mutations in genes encoding ion channel subunits, *KCNQ2* and *SCN2A* (Table 1; Supplementary Materials, Fig. S1A and B). The *KCNQ2* mutation, NM_004518:c.C827T:p.T276I, falls in the highly conserved fifth transmembrane segment of the channel that forms part of the pore. It is two amino acids away from the T274M mutation recently described in an OS patient (26). The *SCN2A* mutation in Patient 3, NM_001040143:c.A5558G:p.H1853R, is in the cytosolic C-terminal region of the protein. It falls within the 250 residue domain that binds FGF14, which is required for localization at the axon initial segment (27). Other *de novo* mutations in the cytosolic domains were recently reported in patients with OS (18). These reports provide strong supporting evidence that these *de novo* mutations, which have not been previously reported in any epilepsy patients, are responsible for OS in these children.

Table 1. Phenotypes and presumed causal mutations in the six trios sequenced

Trio	Phenotype	Age of seizure onset	Current age	Family history	Previous genetic tests	Brain MRI	EEG	Presumed causal mutation	Evidence for pathogenicity
1	OS; severe DD	1 day	5 years	No	arrayCGH, <i>FRAXA</i> , <i>STXBP1</i> , <i>MECP2</i> , <i>CDKL5</i> , <i>POLG</i> , <i>ARX</i>	Age 14 days: reduced posterior white matter volume; thin corpus callosum	Age 14 days: Burst suppression	<i>de novo</i> in <i>KCNQ2</i> : NM_004518:c.C827T:p.T276I	<i>KCNQ2</i> previously implicated in OS
2	OS; metopic synostosis; severe DD	1 day	4 years	Paternal great-grandmother and her sister had epilepsy	arrayCGH, <i>FRAXA</i> , <i>MECP2</i> , <i>CDKL5</i> , <i>STXBP1</i>	Age 1 year: cerebral atrophy with delayed myelination and hypomyelination	Age 14 days: Burst suppression	Recessive variant in <i>KCNT1</i> , homozygous due to UPD9: NM_020822:c.G2896A:p.A966T	<i>KCNT1</i> previously implicated in MMPSI and ADNFLE; electrophysiology demonstrated effect on channel current
3	OS; severe DD	14 days	5 years	No	arrayCGH, <i>CDKL5</i> , <i>ARX</i> , <i>STXBP1</i>	Age 8 months: cerebral atrophy, delayed & reduced myelination	Age 6 weeks: Burst suppression	<i>de novo</i> in <i>SCN2A</i> : NM_001040143:c.A5558G:p.H1853R	<i>SCN2A</i> previously implicated in OS
4	OS; severe DD	4 weeks	Deceased age 2 years, 4 months	Mother's cousin died of seizures at age 2	arrayCGH, <i>MECP2</i> , <i>ARX</i> , <i>STXBP1</i>	Age 9 months: delayed and reduced myelination	Age 3 months: Burst suppression	Simple recessive in <i>PIGQ</i> : NM_004204:exon3:c.690-2A>G	Binding partner <i>PIGA</i> implicated in similar syndrome; mutation leads to exon skipping and reduced GPI synthesis
5	Severe NSEOE; microcephaly; severe DD	2 days	19 years	No	arrayCGH, <i>MECP2</i> , <i>UBE3A</i> , <i>TCF4</i>	Microcephalic (OFC <3rd percentile); structurally normal brain	Age 8 years: multifocal seizure potential on background of significant disruption of cortical function	<i>de novo</i> in <i>CSNK1G1</i> : NM_022048:c.C688T:p.R230W	<i>CSNK1G1</i> involved in synaptic transmission
6	Severe NSEOE; severe DD; PDA and ASD as neonate; cutaneous hypopigmentation	2.5 months	11 years	No	arrayCGH, <i>MECP2</i> , <i>CDKL5</i> , <i>STXBP1</i>	Cerebral hypoplasia; microcephaly (OFC < 0.4th percentile)	Age 6 years: background diffusely of low amplitude, with multifocal sharp waves	<i>de novo</i> in <i>CBL</i> : NM_005188:exon9:c.1228-1G>A	<i>CBL</i> implicated in NCFC syndrome; mutation leads to exon skipping

OS, Ohtahara syndrome; NSEOE, non-syndromic early-onset epilepsy; DD, developmental delay; PDA, patent ductus arteriosus; ASD, atrial septal defect; OFC, occipital frontal cortex; MMPSI, malignant migrating partial seizures of infancy; ADNFLE, autosomal dominant nocturnal frontal lobe epilepsy; GPI, glycosylphosphatidylinositol; NCFC, neuro-cardio-facial-cutaneous; UPD, uniparental disomy. More detailed clinical descriptions, including seizure types, head circumference and treatments, are given in Supplementary Materials, Note S1.

Patient 2: *KCNT1*

Patient 2 had very severe early-onset epilepsy, an EEG consistent with OS (Supplementary Material, Fig. S2), and profound developmental delay. He had a paternal family history of childhood idiopathic epilepsy affecting his father's maternal aunt, grandmother and nephew. Patient 2 did not have any compelling *de novo* mutations. However, low chromosomal heterozygosity and detection of multiple Mendelian errors (Fig. 1; Supplementary Materials, Fig. S1C) suggested that he had paternal

isodisomy for chromosome 9. This was subsequently confirmed by SNP array (Supplementary Materials, Fig. S3; see Supplementary Materials, Note S3). This finding prompted two new alternative hypotheses: that the patient's symptoms were due to aberrant expression of an imprinted gene on chromosome 9, or that there was a recessive pathogenic mutation on this chromosome that had become homozygous as a result of the isodisomy. There was no evidence in the literature for imprinted genes on this chromosome that had plausible links to epilepsy. We therefore scanned the patient's chromosome 9 for rare homozygous

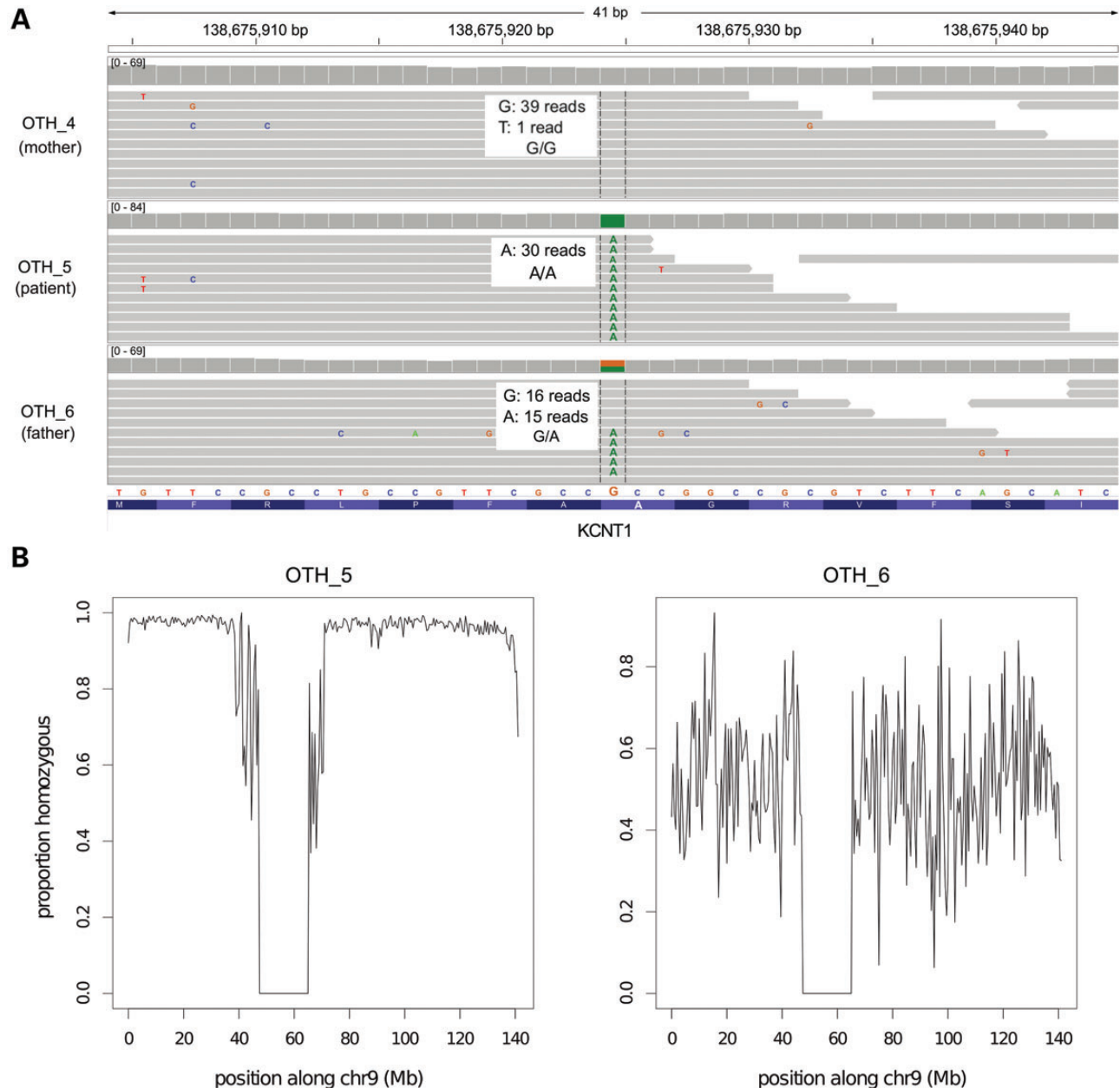


Figure 1. Paternal isodisomy in Patient 2. (A) We observed multiple Mendelian errors on chromosome 9 which led us to suspect uniparental disomy (UPD). All variants in the patient (OTH_5) appeared to have been inherited from his father (OTH_6). The *KCNT1* variant is illustrated here as an example. Grey bars represent individual sequencing reads from the sample indicated on the left, and colored letters divergences from the reference sequence. The grey 'pile-up' along the top indicates the sequence coverage. The genotype of each individual is shown. (B) These plots show the proportion of variants that are homozygous in 500 kb windows across chr9. OTH_5 was completely homozygous, apart from a few spurious calls; the pattern is similar to that seen on chromosome X in males. His father, OTH_6, is shown for comparison. Note that the dip in the middle represents the centromere.

variants that might be pathogenic, including around *STXBPI* and *SPTAN1*. The only plausible candidate was a novel non-synonymous variant in *KCNT1* at 9q34.3 that disrupted a highly conserved alanine residue in the intracellular C-terminal domain, NM_020822:c.G2896A:p.A966T. This gene encodes the Na⁺-activated K⁺ channel known as 'Slack', which is very widely expressed throughout the brain (28). Dominant mutations in *KCNT1* were recently reported to cause autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE) (29), malignant migrating partial seizures of infancy (MMPSI) (30,31) and infantile spasms (32). Interestingly, one of the MMPSI patients with a *KCNT1* mutation was described as having a 'subtle' burst-suppression EEG (30). Distinct from this patient, however, our Patient 2 did not have migrating seizures, and had a clear burst-suppression EEG pattern. Thus, different mutations in *KCNT1* have heterogeneous phenotypic consequences.

The A966 residue in *KCNT1* is completely conserved across all vertebrates for which genome sequences are available. The sequence of the Slack channel C-terminal region is highly conserved between rat and human (92% identical), and residue A966 in the human Slack channel corresponds to residue A945 in the rat protein. We therefore used a rat mutant construct to explore the effect of the novel mutation on channel function, in the same way we described previously for the MMPSI-linked R428Q and A934T mutations (31). We expressed wild-type (WT) and A945T rat Slack in *Xenopus laevis* oocytes, and

measured channel activity by performing two-electrode voltage clamping experiments. These experiments showed that activity of the A945T mutant channel was increased significantly by a factor of 13 relative to the WT at +60 mV (Fig. 2A and B). The amounts and integrity of the WT and A945T cRNA used in these experiments appeared similar, just 8% higher in the mutant than the WT (Fig. 2C).

Slack channel activity increases with depolarization (28,33). We therefore compared the voltage dependence of A945T and WT channels. Channel activity of the A945T mutant was significantly greater than that of the WT at all positive potentials (Fig. 2D). However, the voltage dependence of activation of the A945T mutant with depolarization to positive potentials did not differ from that of the WT channel (Fig. 2D) or from that of the previously published reports of Slack channel voltage dependence (31,33). Together, these results suggest that channel opening probability may be greater in the A945T than the WT channel over the same range of depolarized membrane potentials, which would account for the epileptic activity seen in this patient.

Patient 4: *PIGQ*

Patient 4, who was of West African origin, had severe early-onset epilepsy with a burst-suppression EEG, consistent with OS. Although he was reported to be non-consanguineous,

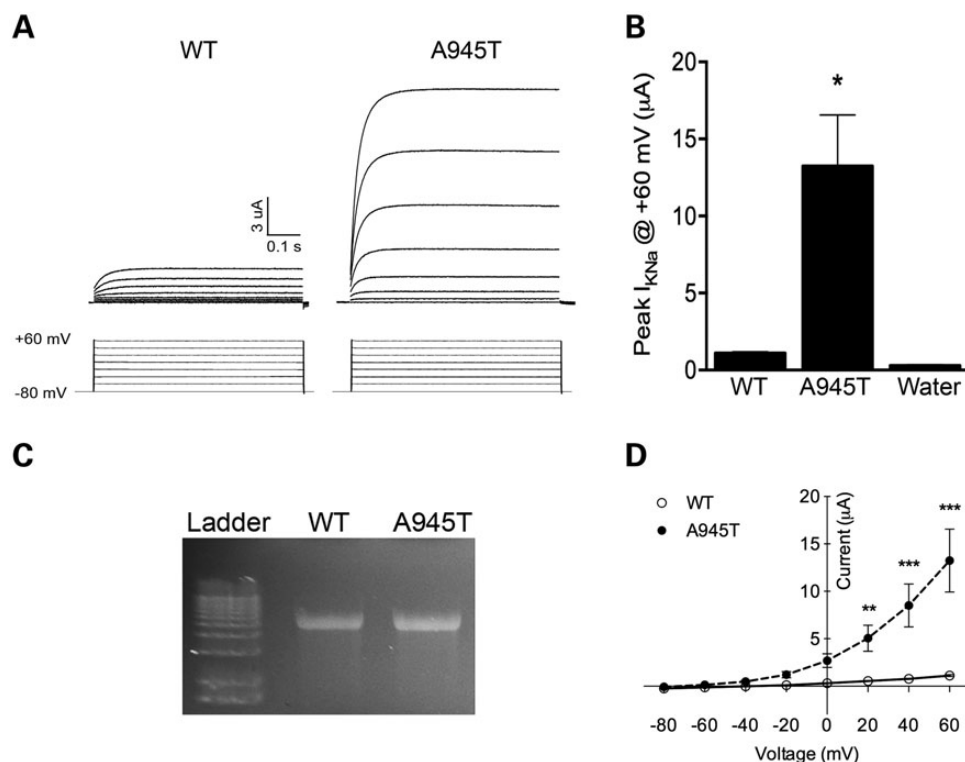


Figure 2. Electrophysiological and channel expression analysis of *KCNT1* mutation found in Patient 2. WT or A945T mutant Slack channel was expressed in *Xenopus laevis* oocytes, and two-electrode voltage clamping (TEVC) performed. (A) A representative trace of current activity recorded from an oocyte expressing WT or A945T is shown on top, with the voltage-clamping protocol displayed underneath. (B) Averaged quantitation of the peak current is compared at +60 mV ($P < 0.001$, $n = 5$, Student's t -test; representative of three independent experiments). (C) The quality of RNA injected into *Xenopus* oocytes was checked on a 1% formaldehyde agarose gel. (D) Current–voltage relations for the WT or A945T channels. Channel activity as measured at peak current amplitude and normalized to the value at +60 mV is plotted against voltage (** $P < 0.01$, *** $P < 0.001$, $n = 5$, two-way ANOVA, Sidak's multiple comparisons test).

we found several extended homozygous regions in his genome. Within a 2 Mb homozygous region on chr16, he had a novel homozygous single nucleotide variant (SNV) that was predicted to disrupt the highly conserved splice acceptor site of exon 3 of the *PIGQ* gene: NM_004204:exon3:c.690–2A>G. Both parents were heterozygous and two unaffected siblings were either heterozygous or homozygous for the reference allele (Supplementary Materials, Fig. S1D). *PIGQ*, formerly called *GPII*, encodes a subunit of an *N*-acetylglucosaminyltransferase that catalyzes the first step in glycosylphosphatidylinositol (GPI) biosynthesis. *PIGQ* stabilizes the enzyme complex (34). A nonsense mutation in the X-linked *PIGA* gene, which encodes another subunit of this enzyme, was recently reported to cause a lethal disorder characterized by multiple congenital abnormalities, structural brain malformations, joint contractures and neonatal seizures with a burst-suppression EEG (35). Recessive mutations in other GPI synthesis genes cause clinically heterogeneous syndromes, all of which involve seizures (36–39). Thus, this *PIGQ* mutation seemed a very plausible candidate for causing OS in Patient 4.

The patient died before we discovered this mutation so we were unable to obtain samples to test the effect of this homozygous mutation on *PIGQ* activity. However, we obtained RNA samples from his parents' blood and examined *PIGQ* splicing. There were two *PIGQ* transcripts, one consistent with the reference annotation and another with a deletion of exon 3, as expected given that the parents were heterozygous for the variant at the splice acceptor site for this exon (Fig. 3A; Supplementary Materials, Fig. S4). Since exon 3 falls immediately before the catalytic domain of *PIGQ*, the mutation seemed likely to abrogate the function of the enzyme and lead to a reduction in GPI synthesis, as was seen for the nonsense mutation in *PIGA* (35). We tested the parents for abnormalities in serum alkaline phosphatase levels and in expression of CD59 on red blood cells, which are typical signs of impaired GPI synthesis (35,36,40), but found none. This is not entirely surprising since heterozygous carriers of other *PIG* gene mutations had normal CD59 levels (37).

The skipping of exon 3 causes an in-frame deletion of 44 amino acids (Fig. 3A). To assess whether this abnormality affected *PIGQ* function *in vitro*, we transfected human *PIGQ* cDNA either with or without exon 3 into *PIGQ*-deficient Chinese hamster ovary (CHO) cell lines. The mutant *PIGQ* did not restore the surface expression of GPI-anchored proteins (GPI-AP) as efficiently as the WT (Fig. 3B). Additionally, the expression of mutant protein was greatly decreased to a level undetectable by western blotting (Fig. 3C). These results demonstrated that the *PIGQ* protein lacking the 44 amino acids had some functional activity but was unstable, and so GPI synthesis was impaired. Over 150 proteins have been reported to have GPI anchors (41), including several with important roles in neural development and function (42–44). Further work is needed to determine which of these provides the causal link with epilepsy.

Patient 5: *CSNK1G1*

Patient 5 had severe tonic-clonic epilepsy, microcephaly and developmental delay. We found a *de novo* non-synonymous mutation at a highly conserved site in *CSNK1G1* (NM_022048: c.C688T:p.R230W; Supplementary Materials, Fig. S1E). This

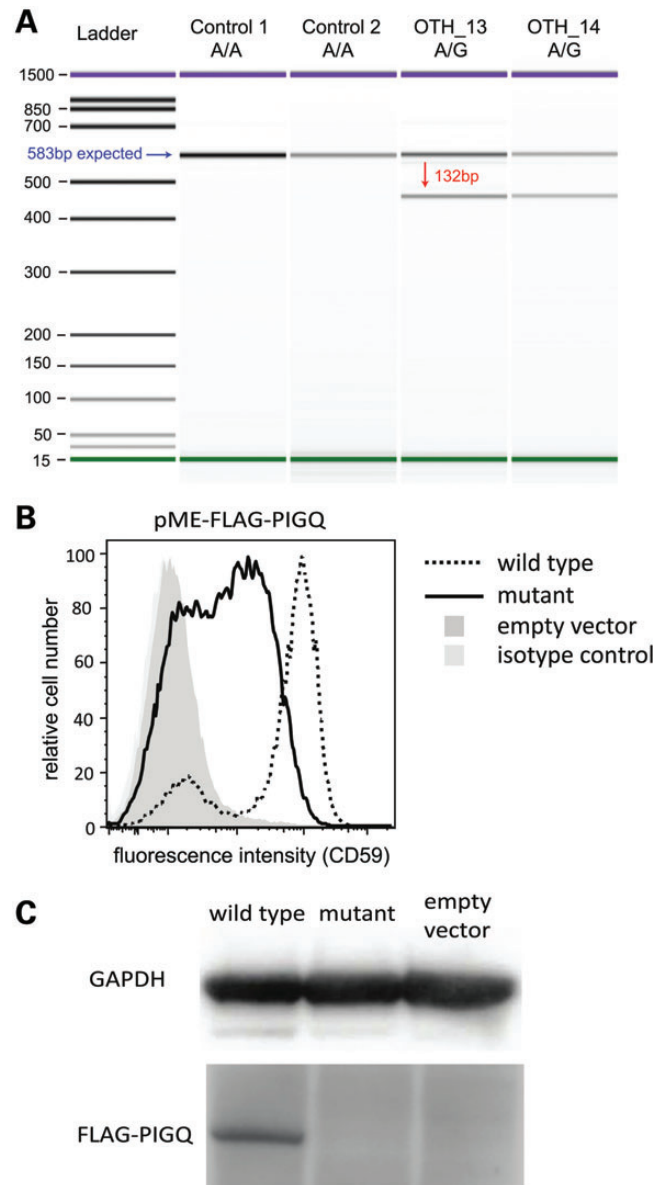


Figure 3. *PIGQ* splicing mutation in Patient 4. (A) The variant causes skipping of exon 3. This image shows the Bioanalyzer gel from an RT-PCR (see Materials and Methods) and demonstrates the presence of two *PIGQ* transcripts in the heterozygous parents (OTH_13, OTH_14). The blue arrow indicates the band expected from the annotated transcript, and the red arrow that expected from the skipping of exon 3. (B) Severely decreased functional activity of the mutant *PIGQ*. *PIGQ*-deficient CHO cells were transiently transfected with WT or mutant *PIGQ* cDNA (lacking exon 3). Restoration of the surface expression of CD59, a GPI-anchored protein, was assessed by flow cytometry after staining with anti-CD59 antibody. The mutant *PIGQ* did not restore the surface expression of CD59 as efficiently as the WT. X axis, fluorescence intensity corresponding to CD59 expression level per cell; Y axis, relative cell number. (C) The expression of mutant protein was greatly decreased and could not be detected by western blotting.

gene encodes casein kinase 1 (CK1), gamma 1, a serine/threonine kinase expressed in many tissues including the brain. CK1 regulates the phosphorylation of NMDA receptors and is thus important for synaptic transmission (45) (see Supplementary Materials, Note S4). The *CSNK1G1* mutation is a prime

candidate in this patient, but genetic validation studies did not find any additional patients (see below) and further work is needed to definitively establish it as causal.

Patient 6: *CBL*

Patient 6 had severe tonic epilepsy, microcephaly and developmental delay. We detected a *de novo* mutation in a highly conserved splice site in the *CBL* gene (NM_005188:exon9: c.1228-1G>A), and showed that this led to exon skipping (Supplementary Materials, Fig. S5). Cbl is a ubiquitously expressed adaptor molecule and ubiquitin ligase that regulates the Ras/MAPK pathway (46). It is primarily recognized as a tumor suppressor (47), but germline mutations in it and other genes involved in the Ras/MAPK pathway can also cause various developmental disorders collectively known as the neuro-cardio-facial-cutaneous (NCFC) syndromes or 'RAS-opathies' (48). Notably, mutations in *KRAS* and *BRAF* were recently reported in two boys with refractory epilepsy and cardio-facial-cutaneous (CFC) syndrome (49). It is likely that this splicing mutation ablates Cbl's ubiquitin ligase activity, thereby over-stimulating Ras/MAPK signaling. This may have disrupted neuronal development and led to severe epilepsy (more details in Supplementary Materials, Note S4).

The patient was reviewed by a number of clinical geneticists. Hypopigmented skin lesions and a history of congenital heart disease were noted, but the clinical diagnosis of a NCFC was not considered likely. Review after this molecular finding did not affect the clinical diagnosis and this patient is still thought not to fit phenotypically into this group of disorders. If this gene is confirmed to be causal, this will widen further the phenotype of the Ras-MAPK disorders.

Mutation screening in other cases

Using a targeted resequencing approach (50), we screened *KCNT1*, *PIGQ*, *CBL* and *CSNK1G1* in a large cohort of epileptic encephalopathy cases from Australia (Supplementary Materials, Table S1). This included two cases of OS, five of early myoclonic encephalopathy [EME; a syndrome which shares features with OS but which is predominantly myoclonic in nature (51)], and thirty-eight of early-onset epileptic encephalopathy (defined as onset within the first three months of life). We also screened these genes in thirteen other cases of OS and EME (Supplementary Materials, Table S2) from a UK cohort, using Sanger sequencing. We looked for coding variants that would fit a recessive model in *KCNT1* and *PIGQ*, or a *de novo* model in *CBL* and *CSNK1G1*, as was observed in our patients, but found none. Two of the OS cases have subsequently been attributed to mutations in other genes (50). Thus, our failure to replicate our findings likely reflects further genetic heterogeneity in severe early-onset epilepsy. We also appreciate the relatively small number of patients we have screened who had similar phenotypes to those described here (total sixteen OS, six EME).

DISCUSSION

WGS heralds promise as a tool for clinical diagnosis of patients with genetic disease. As part of a wide-ranging program to

evaluate the clinical utility of WGS (WGS500), we sequenced six patients with severe early-onset epilepsy who had evaded molecular diagnosis by conventional single-gene clinical screening. In doing so, we identified two new genes for OS, as well as two putative genes for severe early-onset epilepsy. This increases the number of known genes for OS from six [*STXBPI* (14), *ARX* (13), *CDKL5* (16), *KCNQ2* (26), *SCN2A* (18) and *SLC25A22* (52)] to eight (adding *KCNT1* and *PIGQ*). In two cases, we found an unexpected inheritance mechanism: uniparental disomy of chromosome 9 in Patient 2, and simple recessive inheritance due to likely distant consanguinity in Patient 4.

Of particular interest was the discovery that Patient 2 had a pathogenic mutation in *KCNT1* that became homozygous through isodisomy. To date, there have been two published cases of homozygous mutations due to isodisomy causing syndromes involving seizures (53,54), but no examples of such mutations causing severe epilepsy. This is also the first reported case of paternal isodisomy for chromosome 9, and of an apparently recessive mutation in *KCNT1*; recently reported mutations causing MMPSI (31), ADNFLE (29) and infantile spasms (55) were all dominant. Nevertheless, we note that, while this mutation appears to be acting in a recessive manner, there was a paternal history of mild idiopathic epilepsy. It is possible that, in the heterozygous state, this variant predisposes to milder epilepsy, but further testing of affected family members was not possible.

The Slack *KCNT1* A945T mutation had a gain-of-function effect on channel activity. Although the two previously characterized MMPSI-causing mutations, R409Q and A913T, were also gain-of-function alterations, the A945T mutation appears to have a much more profound effect. Whereas channel activity was increased by a factor of 3 at +60 mV in the two MMPSI mutant channels (31), the corresponding increase was 13-fold for the A945T mutation. This observation raises the possibility that the patterns of neuronal firing that produce the distinctive EEG patterns characteristic of these different disorders can vary as a function of *KCNT1* channel activity, and can furthermore influence the nature, severity and onset of the seizures.

The *PIGQ* finding emphasizes the importance of the GPI pathway in brain development and function. This gene was not initially considered as a candidate, since the patient did not have the congenital abnormalities found in children with other *PIG* gene mutations, such as polydactyly. However, it became a very plausible candidate after its binding partner, *PIGA*, was reported to cause a lethal disorder involving neonatal seizures (35). We then demonstrated that the *PIGQ* mutation impairs GPI synthesis in a similar manner (Fig. 3). It remains an open question as to how defective GPI synthesis causes epilepsy, although a number of mechanisms have been suggested, including impaired *Cripto* signaling leading to aberrant forebrain development (44), and disruption of contactin-mediated organization of axonal subdomains at the node of Ranvier (42).

We have also made some unexpected findings about the etiology of other non-syndromic forms of early-onset epilepsy. Although the mutations in patients 5 and 6 have not been definitively established as causal, our results point to several interesting pathways not generally associated with epilepsy. The *de novo* splicing mutation in *CBL* in Patient 6 implies that this patient's condition is actually a 'RAS-opathy' (56), which may be consistent with her congenital heart disease. Other genes in the Ras/MAPK pathway have been reported to cause

epileptic encephalopathy (49) but *CBL* has not. It has been implicated as a cause of juvenile myelomonocytic leukemia (57,58) and also of a Noonan-like syndrome with microcephaly (57,59). A clinical review after our discovery confirmed that Patient 6 did not have facial features typical of Noonan syndrome. Thus, our results suggest that *CBL* mutations may give rise to an even wider spectrum of phenotypes than previously thought.

Our discovery of the *de novo* non-synonymous *CSNK1G1* mutation in Patient 5 hints that the epileptogenic mechanism may involve the Wnt pathway, which CK1 regulates (60,61), although disruptions in synaptic transmission due to abnormal phosphorylation of NMDA receptors (45) would be a more direct explanation. Intriguingly, the *Drosophila* homolog of *CSNK1G3* was found to suppress seizures in the Na⁺-channel gain-of-function mutant *para^{bss1}* (62). Also, a mutation in *PRICKLE1*, which encodes a regulator of the Dishevelled proteins that are intracellular transducers of Wnt signals (63), has been reported to cause progressive myoclonic epilepsy (61). For both *CBL* and *CSNK1G1*, causality can only be definitely established by finding other mutations in patients with similar phenotypes or by extensive functional work in model organisms. Identifying and screening cases similar to Patient 6, rather than those with a more typical NCFC clinical presentation, may increase the chance of finding further patients with *CBL* mutations.

Other groups have already demonstrated the power of WES and WGS in rapidly pinpointing novel genes underlying a rare disease (1,64), particularly in the case of *de novo* inheritance (65). Although all the cases described in this article could probably have been solved by WES, which would have been considerably cheaper, there is emerging evidence to suggest that WES misses clinically relevant mutations because of unequal or incomplete coverage of exons, particularly around the exon boundaries (66). Given that two of our mutations were in splice sites, this was especially relevant. Additionally, the ability to check for pathogenic non-coding mutations in WGS data, as we have done around known early-onset epilepsy genes, is an additional benefit of this approach.

Our study underlines the significant potential of WGS for providing rapid clinical diagnosis of patients with heterogeneous genetic diseases. The patients being investigated here had undergone numerous genetic, biochemical and imaging tests over many years but had been refractory to diagnosis. Using WGS, three of the six patients (Patients 1–3) received a confirmed molecular diagnosis in a clinically relevant timeframe. Conventional molecular testing would not have included these genes. *KCNT1* had not previously been implicated in OS and would not therefore have been tested for this specific phenotype, even though it was described for other severe epilepsy phenotypes after we started this project (29,31). Similarly, *KCNQ2* and *SCN2A* had only been described for benign seizures (67,68) until recent reports of association with the more severe OS phenotype (17,18). These results have already improved the clinical management of these patients' families by providing informed and accurate reproductive risk counseling and the prospect of prenatal diagnosis for future pregnancies.

For the remaining three patients, candidate mutations likely to cause their epilepsy have been identified. The evidence is particularly strong for *PIGQ*, since a homozygous nonsense mutation in *PIGA* causes a similar phenotype (35), and we

demonstrated that our mutation was loss of function. Further genetic and functional validation work is required to prove causality definitively, and this remains a challenge in a clinical setting, particularly for rare diseases. Nevertheless, this situation is expected to improve with greater adoption of these technologies and increased sharing of genetic data in public databases.

In conclusion, our results have led to identification of novel genes for severe epilepsy phenotypes and, in addition, demonstrate the clinical utility of WGS as a means of providing comprehensive and rapid molecular diagnosis for patients with mechanistically complex genetic diseases, with concomitant implications for clinical management of these disorders.

MATERIALS AND METHODS

Description of patients

The six patients were recruited through the Oxford Clinical Genetics department. A summary of the main clinical features is given in Table 1, along with a list of the genetic tests they had undergone before entering this study (all of which were negative). All patients also had extensive metabolic tests on blood, urine and cerebral spinal fluid, all of which were normal. A more detailed clinical description is given in Supplementary Materials, Note S1.

Read mapping and variant calling

WGS was conducted on the Illumina HiSeq platform to a coverage of at least 25×. The reads were mapped to the human reference genome (build 37d5) with Stampy (69), and SNVs and small indels were called with an in-house algorithm, Platypus (70). Variants were annotated relative to RefSeq transcripts using ANNOVAR (71) and relative to all Ensembl transcripts using an in-house tool called VariantAnno.

Variant filtering strategy

To identify *de novo* mutations in the trios, we first screened for variants that were called as homozygous reference in the parents but heterozygous in the child. We then filtered these based on the genotype likelihood ratio (the difference between the log likelihoods for the most likely and the second most likely genotypes), requiring this to be below –5 in all three individuals. This left an average of 126 candidate *de novo* mutations in each trio [about 70 being expected given a mutation rate of 1.18×10^{-8} per base pair per generation (72)]. To remove those likely to be due to technical artifacts or incorrect calling of parental genotypes, we removed variants that had been seen before in WGS500, the 1000 Genomes Project or the NHLBI Exome Sequencing Project (ESP). We then prioritized variants that were predicted to alter the protein sequence in any transcript (non-synonymous SNVs, stop loss or gain variants, indels or splice site mutations), particularly those that were highly conserved across the 46 vertebrate species in the UCSC conservation track. There were 0–3 candidate *de novo* coding mutations per trio.

We also considered a simple recessive or compound heterozygous model in all families, as well as an X-linked model where appropriate. For the compound heterozygous model, we required

two coding variants in the same gene, one inherited from each heterozygous parent. Since EOEE is extremely rare (frequency $\sim 1/100\,000$), we excluded variants with a frequency greater than 0.005 in 1000 Genomes or ESP, or for which there were any homozygotes or hemizygotes or more than five heterozygotes among the other WGS500 samples ($n = 294$), which included no other patients with seizures.

In addition to screening for coding variants, we also looked for variants that might be affecting regulation of known EOEE genes. Specifically, we focused on variants at conserved positions in regulatory regions within 50 kb of the candidate genes *KCNQ2*, *SCN2A*, *SCN1A*, *SPTAN1*, *SRGAP2*, *MAGII*, *PLCB1*, *STXBP1*, *PNPO*, *PCDH19*, *GRIN2A*, *MAPK10*, *CDKL5*, *SLC25A22*, *ERBB4* and *ARX*. A variant was considered conserved if it had a GERP (73) or phyloP (74) score greater than 2 or a phastCons (75) score greater than 0.95, or was in a GERP constrained element (73) or a phastCons constrained element (75). We used the regulatory regions defined by the Ensembl V65 Regulatory Build (http://www.ensembl.org/info/genome/funcgen/regulatory_build.html, last accessed date on February, 2013).

Variant validation

All putatively causal variants were Sanger-sequenced to confirm the genotypes of the proband and parents.

Splicing assays

Fresh blood was collected using PAXgene blood RNA tubes (Becton Dickinson, Oxford, UK) and RNA was extracted using the PAXgene Blood RNA kit (Qiagen). cDNA was synthesized using the QuantiTect Reverse Transcription kit (Qiagen) according to the manufacturer's instructions. We carried out a PCR using the FastStart Kit (Roche) and primers as follows: *PIGQ*-2F, CACGCAGTGAGGTGCTCTT; *PIGQ*-5R, GGGGACATGAGGTGGATGTA; *CBL*-8F, GAGATGGGCTCCAATTCC; *CBL*-11R, GAACTTGGGGCAGATACTGG. To size PCR products accurately, we used 'on-chip-electrophoresis' and ran 1 μ l on a DNA 1000 v2.3 chip using the 2100 Bioanalyzer system (Agilent). We expected WT RT-PCR products of 583 and 698 bp for NM_004204 (*PIGQ*) and NM_005188 (*CBL*), respectively. Sanger sequencing was used to confirm the identity of the aberrant bands.

KCNT1 functional work

Site-directed mutagenesis and cRNA synthesis

The rat homologue (A945T) of the human *KCNT1* A966T mutation was created using the QuikChange Mutagenesis kit (Stratagene) using the WT *KCNT1* construct in pOX expression vector as the template, and the following primers: 5'-GTTCCGCCTGCCATTTGCTACTGGTTCGGGTGTTTAGTA-3' (forward) and 5'-TACTAAACACCCGACCAGTAGCAAATGGCAGGCCGAAC-3' (reverse). The resulting construct was sequenced to confirm the presence of mutation. The cDNA construct was then linearized using NotI, and the complementary RNA (cRNA) made using the mMessage mMachine T3 kit (Ambion). The final reaction was purified using MinElute PCR Purification kit (Qiagen), and RNA eluted in nuclease-free

water. RNA purity and concentration was checked using a NanoDrop reader. RNA quality was also checked on a 1% formaldehyde agarose gel, and its densitometry quantitated using ImageJ to confirm its concentration. cRNA was stored at -20°C until ready for use.

Electrophysiological characterization in *Xenopus laevis* oocytes

All animal procedures were approved by the IACUC at Yale University. Oocytes were prepared as described previously (31). Defolliculated oocytes were injected with 10 ng of RNA encoding WT or A966T, or with sterile water. Oocytes were kept at 18°C , and two-electrode voltage clamping was performed on days 2–5 post-injection, as reported previously (31,76).

PIGQ functional work

We used a previously described human *PIGQ* expression vector (77). The *PIGQ* mutant construct lacking exon 3 was made by site-directed mutagenesis. *PIGQ*-deficient CHO cells (10.2.1) (78) were transiently transfected with FLAG-tagged WT or mutant *PIGQ* cDNA, driven by an SR α promoter (pME FLAG-*PIGQ*) together with a luciferase expression plasmid for monitoring transfection efficiency. Two days later, to determine the restoration of GPI-AP expression, cells were stained with anti-CD59 (5H8) antibody followed by phycoerythrin-conjugated anti-mouse IgG, and analyzed by a flow cytometer (Cant II; BD Biosciences, Franklin Lakes, NJ, USA) using the Flowjo software (Tommy Digital Inc., Tokyo, Japan). To determine FLAG-tagged *PIGQ* protein levels, lysates of transfected cells were subjected to SDS-PAGE, and western blotting was performed using anti-FLAG antibody (M2, Sigma, St Louis, MO, USA), with anti-GAPDH (6C5, Life Technologies, Carlsbad, CA, USA) as a loading control. Transfection efficiency was determined by luciferase activity using the Luciferase assay kit (Promega, Madison, WI, USA).

Screening candidate genes in Australian cohort

We sequenced *PIGQ*, *CSNK1G1*, *CBL* and *KCNT1* in a cohort of 500 epileptic encephalopathy patients from Australia. These patients had a variety of different phenotypes, the breakdown of which is shown in Supplementary Materials, Table S1, with an age of onset ranging from 1 day to 25 years.

We used Molecular Inversion Probes (MIPs) to capture all exon and intron/exon boundaries (5 bp flanking) of target genes (Refseq, hg19 build). Detailed methodology is described elsewhere (79). Briefly, pooled MIPs (Supplementary Materials, Table S3) were used to capture target exons from 100 ng of each proband's DNA and target enrichment was performed by PCR using unique reverse primers for each DNA sample. Pooled libraries were subject to massively parallel sequencing using a 101 paired-end protocol on a HiSeq.

We performed raw read processing as described (79), but use a modified analysis pipeline for variant calling. SNV and indel calling and filtering was performed using the Genome Analysis Tool Kit (GATK version 2.2) (<http://www.broadinstitute.org/gatk/>, last accessed date on February, 2013). We excluded from further analysis any variants with allele balance >0.70 ,

QUAL < 30, QD < 5 or coverage < 25 ×, and variants in clusters (window size 10 bp) or in homopolymer runs (5 bp). Variants were annotated with SeattleSeq (version 134; <http://snpgs.washington.edu/SeattleSeqAnnotation134/>) and the Exome Sequencing Project dataset (see <http://eversusgs.washington.edu/EVS/>, last accessed date on February, 2013) used to assess variant frequency in the control population. For dominant (or *de novo*) models, we considered only variants not present in this control sample set. For recessive candidates, we considered variants with a frequency in controls of < 1% (European American control frequency). Only non-synonymous, splice-site or frameshift variants were assessed further.

Where family members were available, segregation analysis was performed using a 'MIP-pick' strategy. We selected and re-pooled only the MIPs that captured the genomic sequence harboring the rare variant of interest and performed target enrichment PCR and sequencing as above for all relevant probands and family members.

Screening candidate genes in UK cases

We carried out Sanger sequencing of *CBL*, *CSNK1G1*, *PIGQ* and *KCNT1* in 11 patients with Ohtahara syndrome. Clinical details are given in Supplementary Materials, Table S2. Genomic DNA from participating individuals was extracted from peripheral lymphocytes by standard techniques. All participants gave written informed consent and the study was performed in accordance with the Declaration of Helsinki.

Primer pairs were designed for all coding exons with primer3 software (1,2) (<http://bioinfo.ut.ee/primer3/>, last accessed date on February, 2013) (Supplementary Materials, Table S4). The exons were amplified by PCR using BioMix™ Red (Bioline Ltd). Two different PCR conditions were carried out to amplify exons: (i) an initial denaturation of 95°C for 5 min, followed by 35 cycles of 45 s denaturation at 95°C, 45 s annealing at 58–62°C (depending on fragment) and 1 min extension at 72°C with a final extension at 72°C for 5 min, or (ii) a touchdown PCR program: an initial denaturation of 95°C for 5 min, followed by 24 cycles of 30 s denaturation at 95°C, 30 s annealing at 62°C (minus 0.5°C per cycle) and 1 min extension at 72°C, followed by 15 cycles of 30 s denaturation at 95°C, 30 s annealing at 50°C and 1 min extension at 72°C with a final extension at 72°C for 10 min. If PCR condition 1 was not successful, PCR condition 2 was applied. PCR products were cleaned up with MicroCLEAN (Web Scientific) and were directly sequenced by the Big Dye Terminator Cycle Sequencing System (Applied Biosystems Inc.). Sequencing reactions were run on an ABI PRISM 3730 DNA Analyzer (Applied Biosystems Inc.) and analyzed using Chromas (<http://www.technelysium.com.au/chromas.html>, last accessed date on February, 2013).

AUTHOR CONTRIBUTIONS

H.C.M. analyzed the WGS data. L.K.K. directed the *KCNT1* electrophysiology experiments, which were performed by G.E.K., M.F., M.R.B. and J.K. A.T.P. and K.A.H. did the Sanger sequencing and splicing assays. G.B. made the *KCNT1* construct for electrophysiology, supervised by R.N. J.B., A.K. and J.-B.C. created the bioinformatics infrastructure of the

WGS500 project and provided NGS data processing, preparing the BAM files and variant calls. R.C. and A.R. contributed to WGS analysis. Y.M. and T.K. did the *in vitro* *PIGQ* experiments. S.H. and I.E.S. studied patients in the Australian cohort and performed phenotyping analysis of the larger epilepsy panel, on which G.C. and H.M. performed the MIP sequencing. M.A.K. and E.M. performed phenotyping and Sanger sequencing on the UK cohort. H.S., D.S. and E.B. contributed samples and clinical data from the affected individuals and assisted with the interpretation of results. Z.Z. and T.M. provided clinical data and advice on the phenotypes. E.B., D.B., G.M., J.C.T. and P.D. conceived the study, and L.K.K., E.B., P.D. and J.T. directed it. H.C.M., G.E.K., A.T.P., E.B., I.E.S., L.K.K., J.C.T. and P.D. wrote the article. This project was carried out as part of the WGS500 Consortium.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank the patients and their families for participating in this study, the Genomics Core at the Wellcome Trust Centre for Human Genetics (WTCHG) for generating the WGS data, Elham Sadighi Akha for running the SNP arrays for trio 2 and Kevin Leyden for doing the CD59 flow cytometry assay for trio 4. We are also grateful to Adeline Nogh and Amy McTague for their input in gathering the UCL cohort, and to Helen Cross and Jozef Gecz for facilitating initial contact between groups.

Conflict of Interest statement. D.B. is an employee of Illumina Inc.

FUNDING

This work was supported in part by a Wellcome Trust Core Award (090532/Z/09/Z) to the Wellcome Trust Centre for Human Genetics and a Wellcome Trust Senior Investigator Award to P.D. (095552/2/11/2), in part by NIH grant HD067517 and in part by the Oxford NIHR Biomedical Research Centre. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust and the Oxford NIHR Biomedical Research Centre.

REFERENCES

1. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
2. Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H. and Inoue, I. (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J. Hum. Genet.*, **57**, 621–632.
3. Palles, C., Cazier, J.B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z., Spain, S.L., Almeida, E.G., Salguero, I. *et al.* (2012) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.*, **45**, 136–144.
4. Lise, S., Clarkson, Y., Perkins, E., Kwasniewska, A., Sadighi Akha, E., Schnekenberg, R.P., Suminaite, D., Hope, J., Baker, I., Gregory, L. *et al.*

- (2012) Recessive mutations in SPTBN2 implicate beta-III spectrin in both cognitive and motor development. *PLoS Genet.*, **8**, e1003074.
5. Sharma, V.P., Fenwick, A.L., Brockkop, M.S., McGowan, S.J., Goos, J.A., Hoogeboom, A.J., Brady, A.F., Jeelani, N.O., Lynch, S.A., Mulliken, J.B. *et al.* (2013) Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis. *Nat. Genet.*, **45**, 304–307.
 6. Nordli, D.R. Jr. (2012) Epileptic encephalopathies in infants and children. *J. Clin. Neurophysiol.*, **29**, 420–424.
 7. Kadera, H., Kato, M., Nord, A.S., Walsh, T., Lee, M., Yamanaka, G., Tohyama, J., Nakamura, K., Nakagawa, E., Ikeda, T. *et al.* (2013) Targeted capture and sequencing for detection of mutations causing early onset epileptic encephalopathy. *Epilepsia*, **54**, 1262–1269.
 8. Berg, A.T., Berkovic, S.F., Brodie, M.J., Buchhalter, J., Cross, J.H., van Emde Boas, W., Engel, J., French, J., Glauser, T.A., Mathern, G.W. *et al.* (2010) Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005–2009. *Epilepsia*, **51**, 676–685.
 9. Ohtahara, S. and Yamatogi, Y. (2003) Epileptic encephalopathies in early infancy with suppression-burst. *J. Clin. Neurophysiol.*, **20**, 398–407.
 10. Pavone, P., Spalice, A., Polizzi, A., Parisi, P. and Ruggieri, M. (2012) Ohtahara syndrome with emphasis on recent genetic discovery. *Brain Dev.*, **34**, 459–468.
 11. Hino-Fukuyo, N., Haginoya, K., Iinuma, K., Uematsu, M. and Tsuchiya, S. (2009) Neuroepidemiology of West syndrome and early infantile epileptic encephalopathy in Miyagi Prefecture, Japan. *Epilepsy Res.*, **87**, 299–301.
 12. Tavyev Asher, Y.J. and Scaglia, F. (2012) Molecular bases and clinical spectrum of early infantile epileptic encephalopathies. *Eur. J. Med. Genet.*, **55**, 299–306.
 13. Kato, M., Saitoh, S., Kamei, A., Shiraishi, H., Ueda, Y., Akasaka, M., Tohyama, J., Akasaka, N. and Hayasaka, K. (2007) A longer polyalanine expansion mutation in the ARX gene causes early infantile epileptic encephalopathy with suppression-burst pattern (Ohtahara syndrome). *Am. J. Hum. Genet.*, **81**, 361–366.
 14. Saito, H., Kato, M., Mizuguchi, T., Hamada, K., Osaka, H., Tohyama, J., Urano, K., Kumada, S., Nishiyama, K., Nishimura, A. *et al.* (2008) De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat. Genet.*, **40**, 782–788.
 15. Deprez, L., Weckhuysen, S., Holmgren, P., Suls, A., Van Dyck, T., Goossens, D., Del-Favero, J., Jansen, A., Verhaert, K., Lagae, L. *et al.* (2010) Clinical spectrum of early-onset epileptic encephalopathies associated with STXBP1 mutations. *Neurology*, **75**, 1159–1165.
 16. Melani, F., Mei, D., Pisano, T., Savasta, S., Franzoni, E., Ferrari, A.R., Tatini, C. and Guerrini, R. (2011) CDKL5 gene-related epileptic encephalopathy: electroclinical findings in the first year of life. *Dev. Med. Child Neurol.*, **53**, 354–360.
 17. Saito, H., Kato, M., Koide, A., Goto, T., Fujita, T., Nishiyama, K., Tsurusaki, Y., Doi, H., Miyake, N., Hayasaka, K. *et al.* (2012) Whole exome sequencing identifies KCNQ2 mutations in Ohtahara syndrome. *Ann. Neurol.*, **72**, 298–300.
 18. Nakamura, K., Kato, M., Osaka, H., Yamashita, S., Nakagawa, E., Haginoya, K., Tohyama, J., Okuda, M., Wada, T., Shimakawa, S. *et al.* (2013) Clinical spectrum of SCN2A mutations expanding to Ohtahara syndrome. *Neurology*, **81**, 992–998.
 19. Touma, M., Joshi, M., Connolly, M.C., Grant, P.E., Hansen, A.R., Khwaja, O., Berry, G.T., Kinney, H.C., Poduri, A. and Agrawal, P.B. (2013) Whole genome sequencing identifies SCN2A mutation in monozygotic twins with Ohtahara syndrome and unique neuropathologic findings. *Epilepsia*, **54**, e81–e85.
 20. Veeramah, K.R., O'Brien, J.E., Meisler, M.H., Cheng, X., Dib-Hajj, S.D., Waxman, S.G., Talwar, D., Girirajan, S., Eichler, E.E., Restifo, L.L. *et al.* (2012) De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am. J. Hum. Genet.*, **90**, 502–510.
 21. Molinari, F., Raas-Rothschild, A., Rio, M., Fiermonte, G., Encha-Razavi, F., Palmieri, L., Palmieri, F., Ben-Neriah, Z., Kadhom, N., Vekemans, M. *et al.* (2005) Impaired mitochondrial glutamate transport in autosomal recessive neonatal myoclonic epilepsy. *Am. J. Hum. Genet.*, **76**, 334–339.
 22. Saito, H., Tohyama, J., Kumada, T., Egawa, K., Hamada, K., Okada, I., Mizuguchi, T., Osaka, H., Miyata, R., Furukawa, T. *et al.* (2010) Dominant-negative mutations in alpha-II spectrin cause West syndrome with severe cerebral hypomyelination, spastic quadriplegia, and developmental delay. *Am. J. Hum. Genet.*, **86**, 881–891.
 23. Kurian, M.A., Meyer, E., Vassallo, G., Morgan, N.V., Prakash, N., Pasha, S., Hai, N.A., Shuib, S., Rahman, F., Wassmer, E. *et al.* (2010) Phospholipase C beta 1 deficiency is associated with early-onset epileptic encephalopathy. *Brain*, **133**, 2964–2970.
 24. Shen, J., Gilmore, E.C., Marshall, C.A., Haddadin, M., Reynolds, J.J., Eyaid, W., Bodell, A., Barry, B., Gleason, D., Allen, K. *et al.* (2010) Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nat. Genet.*, **42**, 245–249.
 25. Mills, P.B., Surtees, R.A., Champion, M.P., Beesley, C.E., Dalton, N., Scambler, P.J., Heales, S.J., Briddon, A., Scheimberg, I., Hoffmann, G.F. *et al.* (2005) Neonatal epileptic encephalopathy caused by mutations in the PNPO gene encoding pyridox(am)ine 5'-phosphate oxidase. *Hum. Mol. Genet.*, **14**, 1077–1086.
 26. Weckhuysen, S., Mandelstam, S., Suls, A., Audenaert, D., Deconinck, T., Claes, L.R., Deprez, L., Smets, K., Hristova, D., Yordanova, I. *et al.* (2012) KCNQ2 encephalopathy: emerging phenotype of a neonatal epileptic encephalopathy. *Ann. Neurol.*, **71**, 15–25.
 27. Laezza, F., Lampert, A., Kozel, M.A., Gerber, B.R., Rush, A.M., Nerbonne, J.M., Waxman, S.G., Dib-Hajj, S.D. and Ornitz, D.M. (2009) FGF14 N-terminal splice variants differentially modulate Nav1.2 and Nav1.6-encoded sodium channels. *Mol. Cell. Neurosci.*, **42**, 90–101.
 28. Bhattacharjee, A., Gan, L. and Kaczmarek, L.K. (2002) Localization of the Slack potassium channel in the rat central nervous system. *J. Comp. Neurol.*, **454**, 241–254.
 29. Heron, S.E., Smith, K.R., Bahlo, M., Nobili, L., Kahana, E., Licchetta, L., Oliver, K.L., Mazarib, A., Afawi, Z., Korczyn, A. *et al.* (2012) Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. *Nat. Genet.*, **44**, 1188–1190.
 30. McTague, A., Appleton, R., Avula, S., Cross, J.H., King, M.D., Jacques, T.S., Bhate, S., Cronin, A., Curran, A., Desurkar, A. *et al.* (2013) Migrating partial seizures of infancy: expansion of the electroclinical, radiological and pathological disease spectrum. *Brain*, **136**, 1578–1591.
 31. Barcia, G., Fleming, M.R., Deligniere, A., Gazula, V.R., Brown, M.R., Langouet, M., Chen, H., Kronengold, J., Abhyankar, A., Cilio, R. *et al.* (2012) De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nat. Genet.*, **44**, 1255–1259.
 32. Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., Han, Y. *et al.* (2013) De novo mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.
 33. Joiner, W.J., Tang, M.D., Wang, L.Y., Dworetzky, S.I., Boissard, C.G., Gan, L., Gribkoff, V.K. and Kaczmarek, L.K. (1998) Formation of intermediate-conductance calcium-activated potassium channels by interaction of Slack and Slo subunits. *Nat. Neurosci.*, **1**, 462–469.
 34. Hong, Y., Ohishi, K., Watanabe, R., Endo, Y., Maeda, Y. and Kinoshita, T. (1999) GPII stabilizes an enzyme essential in the first step of glycosylphosphatidylinositol biosynthesis. *J. Biol. Chem.*, **274**, 18582–18588.
 35. Johnston, J.J., Gropman, A.L., Sapp, J.C., Teer, J.K., Martin, J.M., Liu, C.F., Yuan, X., Ye, Z., Cheng, L., Brodsky, R.A. *et al.* (2012) The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. *Am. J. Hum. Genet.*, **90**, 295–300.
 36. Maydan, G., Noyman, I., Har-Zahav, A., Neriah, Z.B., Pasmanik-Chor, M., Yeheskel, A., Albin-Kaplanski, A., Maya, I., Magal, N., Birk, E. *et al.* (2011) Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. *J. Med. Genet.*, **48**, 383–389.
 37. Almeida, A.M., Murakami, Y., Layton, D.M., Hillmen, P., Sellick, G.S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L. *et al.* (2006) Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nat. Med.*, **12**, 846–851.
 38. Krawitz, P.M., Schweiger, M.R., Rodelsperger, C., Marcelis, C., Kolsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S. *et al.* (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat. Genet.*, **42**, 827–829.
 39. Thompson, M.D., Roscioli, T., Marcelis, C., Nezarati, M.M., Stolte-Dijkstra, I., Sharom, F.J., Lu, P., Phillips, J.A., Sweeney, E., Robinson, P.N. *et al.* (2012) Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). *Am. J. Med. Genet. A*, **158A**, 553–558.
 40. Krawitz, P.M., Murakami, Y., Hecht, J., Kruger, U., Holder, S.E., Mortier, G.R., Delle Chiaie, B., De Baere, E., Thompson, M.D., Roscioli, T. *et al.* (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway,

- cause hyperphosphatasia with mental retardation. *Am. J. Hum. Genet.*, **91**, 146–151.
41. Fujita, M. and Kinoshita, T. (2012) GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. *Biochim. Biophys. Acta*, **1821**, 1050–1058.
 42. Labasque, M. and Faivre-Sarrailh, C. (2010) GPI-anchored proteins at the node of Ranvier. *FEBS Lett.*, **584**, 1787–1792.
 43. Borrie, S.C., Baumer, B.E. and Bandtlow, C.E. (2012) The Nogo-66 receptor family in the intact and diseased CNS. *Cell Tissue Res.*, **349**, 105–117.
 44. McKean, D.M. and Niswander, L. (2012) Defects in GPI biosynthesis perturb Crip1 signaling during forebrain development in two new mouse models of holoprosencephaly. *Biol. Open*, **1**, 874–883.
 45. Chergui, K., Svenningsson, P. and Greengard, P. (2005) Physiological role for casein kinase I in glutamatergic synaptic transmission. *J. Neurosci.*, **25**, 6601–6609.
 46. Swaminathan, G. and Tsygankov, A.Y. (2006) The Cbl family proteins: ring leaders in regulation of cell signaling. *J. Cell. Physiol.*, **209**, 21–43.
 47. Kales, S.C., Ryan, P.E., Nau, M.M. and Lipkowitz, S. (2010) Cbl and human myeloid neoplasms: the Cbl oncogene comes of age. *Cancer Res.*, **70**, 4789–4794.
 48. Denayer, E. and Legius, E. (2007) What's new in the neuro-cardio-facial-cutaneous syndromes? *Eur. J. Pediatr.*, **166**, 1091–1098.
 49. Adachi, M., Abe, Y., Aoki, Y. and Matsubara, Y. (2012) Epilepsy in RAS/MAPK syndrome: two cases of cardio-facio-cutaneous syndrome with epileptic encephalopathy and a literature review. *Seizure*, **21**, 55–60.
 50. Carvill, G.L., Heaven, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., Cook, J., Khan, A., Dorschner, M.O., Weaver, M., Calvert, S. *et al.* (2013) Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat. Genet.*, **45**, 825–830.
 51. Djukic, A., Lado, F.A., Shinnar, S. and Moshe, S.L. (2006) Are early myoclonic encephalopathy (EME) and the Ohtahara syndrome (EIEE) independent of each other? *Epilepsy Res.*, **70** (Suppl. 1), S68–S76.
 52. Molinari, F. (2010) Mitochondria and neonatal epileptic encephalopathies with suppression burst. *J. Bioenerg. Biomembr.*, **42**, 467–471.
 53. Vantaggiato, C., Redaelli, F., Falcone, S., Perrotta, C., Tonelli, A., Bondioni, S., Morbin, M., Riva, D., Saletti, V., Bonaglia, M.C. *et al.* (2009) A novel CLN8 mutation in late-infantile-onset neuronal ceroid lipofuscinosis (LINCL) reveals aspects of CLN8 neurobiological function. *Hum. Mutat.*, **30**, 1104–1116.
 54. Gumus, H., Ghesquiere, S., Per, H., Kondolot, M., Ichida, K., Poyrazoglu, G., Kumandas, S., Engelen, J., Dundar, M. and Caglayan, A.O. (2010) Maternal uniparental isodisomy is responsible for serious myoblasten cofactor deficiency. *Dev. Med. Child Neurol.*, **52**, 868–872.
 55. Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., Han, Y. *et al.* (2013) De novo mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.
 56. Zenker, M. (2011) Clinical manifestations of mutations in RAS and related intracellular signal transduction factors. *Curr. Opin. Pediatr.*, **23**, 443–451.
 57. Niemeyer, C.M., Kang, M.W., Shin, D.H., Furlan, I., Erlacher, M., Bunin, N.J., Bunda, S., Finklestein, J.Z., Sakamoto, K.M., Gorr, T.A. *et al.* (2010) Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat. Genet.*, **42**, 794–800.
 58. Perez, B., Mechinaud, F., Galambun, C., Ben Romdhane, N., Isidor, B., Philip, N., Derain-Court, J., Cassinat, B., Lachenaud, J., Kaltenbach, S. *et al.* (2010) Germline mutations of the CBL gene define a new genetic syndrome with predisposition to juvenile myelomonocytic leukaemia. *J. Med. Genet.*, **47**, 686–691.
 59. Martinelli, S., De Luca, A., Stellacci, E., Rossi, C., Checquolo, S., Lepri, F., Caputo, V., Silvano, M., Buscherini, F., Consoli, F. *et al.* (2010) Heterozygous germline mutations in the CBL tumor-suppressor gene cause a Noonan syndrome-like phenotype. *Am. J. Hum. Genet.*, **87**, 250–257.
 60. Fossat, N., Jones, V., Garcia-Garcia, M.J. and Tam, P.P. (2012) Modulation of WNT signaling activity is key to the formation of the embryonic head. *Cell Cycle*, **11**, 26–32.
 61. Bassuk, A.G., Wallace, R.H., Buhr, A., Buller, A.R., Afawi, Z., Shimojo, M., Miyata, S., Chen, S., Gonzalez-Alegre, P., Griesbach, H.L. *et al.* (2008) A homozygous mutation in human PRICKLE1 causes an autosomal-recessive progressive myoclonus epilepsy-ataxia syndrome. *Am. J. Hum. Genet.*, **83**, 572–581.
 62. Howlett, I.C., Rusan, Z.M., Parker, L. and Tanouye, M.A. (2013) Drosophila as a Model for Intractable Epilepsy: Gilgamesh Suppresses Seizures in parabss1 Heterozygote Flies. *G3 (Bethesda)*, **3**, 1399–1407.
 63. Chan, D.W., Chan, C.Y., Yam, J.W., Ching, Y.P. and Ng, I.O. (2006) Prickle-1 negatively regulates Wnt/beta-catenin pathway by promoting Dishevelled ubiquitination/degradation in liver cancer. *Gastroenterology*, **131**, 1218–1227.
 64. Foo, J.N., Liu, J.J. and Tan, E.K. (2012) Whole-genome and whole-exome sequencing in neurological diseases. *Nat. Rev. Nephrol.*, **8**, 508–517.
 65. Veltman, J.A. and Brunner, H.G. (2012) De novo mutations in human genetic disease. *Nat. Rev. Genet.*, **13**, 565–575.
 66. Sikkema-Raddatz, B., Johansson, L.F., de Boer, E.N., Almomani, R., Boven, L.G., van den Berg, M.P., van Spaendonck-Zwarts, K.Y., van Tintelen, J.P., Sijmons, R.H., Jongbloed, J.D. *et al.* (2013) Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Hum. Mutat.*, **34**, 1035–1042.
 67. Singh, N.A., Westenskow, P., Charlier, C., Pappas, C., Leslie, J., Dillon, J., Anderson, V.E., Sanguinetti, M.C. and Leppert, M.F. (2003) KCNQ2 and KCNQ3 potassium channel genes in benign familial neonatal convulsions: expansion of the functional and mutation spectrum. *Brain*, **126**, 2726–2737.
 68. Berkovic, S.F., Heron, S.E., Giordano, L., Marini, C., Guerrini, R., Kaplan, R.E., Gambardella, A., Steinlein, O.K., Grinton, B.E., Dean, J.T. *et al.* (2004) Benign familial neonatal-infantile seizures: characterization of a new sodium channelopathy. *Ann. Neurol.*, **55**, 550–557.
 69. Lunter, G. and Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, **21**, 936–939.
 70. Pagnamenta, A.T., Lise, S., Harrison, V., Stewart, H., Jayawant, S., Quaghebeur, G., Deng, A.T., Murphy, V.E., Akha, E.S., Rimmer, A. *et al.* (2012) Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J. Hum. Genet.*, **57**, 70–72.
 71. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
 72. Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.
 73. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
 74. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
 75. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
 76. Chen, W., Han, Y., Chen, Y. and Astumian, D. (1998) Electric field-induced functional reductions in the K⁺ channels mainly resulted from supramembrane potential-mediated electroconformational changes. *Biophys. J.*, **75**, 196–206.
 77. Watanabe, R., Inoue, N., Westfall, B., Taron, C.H., Orlean, P., Takeda, J. and Kinoshita, T. (1998) The first step of glycosylphosphatidylinositol biosynthesis is mediated by a complex of PIG-A, PIG-H, PIG-C and GPI1. *EMBO J.*, **17**, 877–885.
 78. Ashida, H., Hong, Y., Murakami, Y., Shishioh, N., Sugimoto, N., Kim, Y.U., Maeda, Y. and Kinoshita, T. (2005) Mammalian PIG-X and yeast Pbn1p are the essential components of glycosylphosphatidylinositol-mannosyltransferase I. *Mol. Biol. Cell.*, **16**, 1439–1448.
 79. O'Roak, B.J., Vives, L., Fu, W., Egerton, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K. *et al.* (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**, 1619–1622.