Application of emerging sequencing and genomic technologies to diagnostic molecular genetics

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN PATHOLOGY

HANNAH LOUISE KENNEDY

OTAGO UNIVERSITY,
CHRISTCHURCH, NEW ZEALAND

SAPERE AUDE

2017

Attribution of collaborative contributions

Julia Horsfield, Bryony Leeke, and Megan Leask - Otago Zebrafish Facility, Dunedin, NZ
All zebrafish experiments were performed at the Otago Zebrafish Facility under Julia's guidance. Experimental design was a collaborative effort between Hannah Kennedy, Peter George, Kit Doudney, and the Horsfield team. Morpholino technical work was performed by Hannah, Bryony and Megan. All CRISPR/cas9 experiments were performed by the Horsfield team. Handling of the adult zebrafish was performed by the Horsfield team, Hannah was involved in microinjection and characterisation of zebrafish embryos only due to training regulations.

Thomas Hoefken - Brunel University, Uxbridge, London, UK
Yeast experimental design was led by Thomas Hoefken at Brunel University, Uxbridge, London, UK. Yeast construct design was performed by Hannah. Yeast growth assays of a ppa2 knockout strain and measurements of oxidative stress sensitivity were initially performed by Hannah under the guidance of Thomas. Yeast complementation experiments with wildtype and mutant PPA2 were initially performed by Hannah under the guidance of Thomas, however repeat experimentation with new constructs was performed by Thomas after Hannah had returned to NZ.

Andree Pearson, Mark Hampton - Free Radical Research Group, University of Otago-Christchurch, Christchurch, NZ
All seahorse analyses performed on Family Christchurch were led by Mark and Andree. Andree performed the technical aspects of experimentation.

Johannes Mayr - Paracelsus Medical University, Salzburg, Austria
Western blot analysis and PPA2 enzyme activity experiments in fibroblast mitochondria were led by Johannes Mayr .

Tobias Haack and Holger Prokisch- Helmholtz Zentrum, Munich
Clinical case details on families Northern Island, and Basel were compiled by Tobias Haack and Holger Prokisch.

Robert Taylor- Wellcome Trust Centre for Mitochondrial Research, New-

castle, UK
Clinical case details on family Leeds was provided by Robert Taylor.

Zoe Ward - Canterbury Health Laboratories, Christchurch NZ
Zoe provided Perl scripts for use in Chapter 3.

Abstract

The advent of massively parallel sequencing (MPS) has provided much broader opportunities in the field of molecular genomics. Previously, sequencing approaches were limited to a small number of single genes or exons. MPS has enabled assessment of hundreds or thousands of genes associated with a particular phenotype simultaneously, providing the clinical scientist with more information than ever before. This thesis examines the processes and challenges of introducing whole exome sequencing (WES) technologies into a clinical laboratory in three families, each with a distinct segregation pattern of an undiagnosed clinical condition. The challenges associated with the sequencing, analysis and interpretation of sequence variants, as well as elucidating mechanism(s) of pathogenicity of detected variants, were the focus of this thesis.

One family with five male siblings affected with intellectual disability were initially investigated for a suspected X-linked disorder. Although candidate variants were identified, the results did not ultimately detect any clearly pathogenic variants or therefore a diagnosis.

In a second family, a single affected daughter presented with an undefined ageing disorder. We identified a novel *de novo* X-chromosome variant in the gene BCL-6 co-repressor (BCOR), a G to A transition at c.3907G (p.Gly1303Ser). Although functional studies were not performed and definitive pathogenicity was not established during the course of this work, the BCOR variant was considered a strong candidate for this phenotype. Skewed X-chromosome inactivation was identified in this individual which is suggestive of an X-linked disorder, further supporting the role of BCOR dysfunction in her symptoms.

The bulk of this thesis then focused on the analysis of a third pedigree in which four children have acute alcohol sensitivity and pronounced cardiac fibrosis, two of whom died suddenly. Exome sequencing revealed compounding mutations in the *PPA2* gene, inherited recessively in all four children. *PPA2* had not been previously associated with human disease, however, the enzyme encoded by *PPA2* localises to the mitochondria, which implicated this enzyme in this family's disorder. *In silico* algorithms supported a damaging effect of both variants upon the protein, and early functional experiments in yeast and zebrafish were suggestive that PPA2 was important in cardiac function.

This finding led to the identification of three additional families with *PPA2* mutations [1], in collaboration with laboratories in Germany and Austria. The phenotype of affected individuals in the European families was more severe resulting in infantile death, whereas the NZ family experienced death in the second decade of life following ethanol exposure. This work has established *PPA2* as a new cardiac disease gene, and ended a diagnosis odyssey for a NZ family. It has also instigated an ongoing investigation into the mechanism of PPA2-associated disease.

Exome sequencing in these three families has highlighted both the power of MPS to provide a diagnosis, and the challenges associated with analysing such complex data. In particular, the work has established a pipeline for the clinical laboratory which will need an associated laboratory for extensive follow-up functional analysis of candidate variants where a clinically significant impact has not previously been proved.

Acknowledgements

The process of completing this work has been incredibly rewarding and enjoyable. I owe that to the tremendous support I have received from my family, friends, colleagues, and supervisors.

Professor Peter George initially encouraged me to undertake further study, and was instrumental in getting me started on this PhD. Thank you Peter for your guidance, motivation and support over the past 6 years as my supervisor. You made many things possible for me, and I will be forever grateful for your input into my early career.

My gratitude also goes to my supervisor Dr Kit Doudney. Kit, you have been a colleague and a friend for many years, and I can't think of anyone better to have had on my team. Thank you for the coffee dates and crisis talks, your optimism, and your never ending patience. Your help in preparing this thesis was invaluable.

Thank you also to my colleagues and friends in Molecular Pathology and Canterbury Health Laboratories. Every one of you has contributed in some way to my thesis, and I would like to acknowledge the many useful ideas, discussions, and support you all provided. Special mention needs to go to Heather Barnes, Howard Potter, Zoe Ward and Darrell Wang for helping me to manage my workload through my PhD, Dr Stephen Brennan for encouraging curiosity, Dr Alexa Kidd for her clinical expertise, and Trevor Anderson for giving me a quiet space to write in.

To all my collaborators, in particular Julia Horsfield, Thomas Hoefken, Tobias Haack and Hans Mayr, thank you for sharing your knowledge and skills. Your contribution to this work is much appreciated.

To my family, thank you for raising me in an environment that made me believe a PhD was achievable. Your support and encouragement has been so appreciated, and I promise that I will try to talk less genetics with dad over family dinners in the future.

To Matt, thank you for putting up with the PhD-related drama, stress, despair and excitement that have been a weekly occurrence since you met me. I won't make any promises that this will change.

Perhaps the biggest thanks of all need to go to the subjects of this study. In particular John, Margaret, Hannah, and Stephen, and even in their absence, Gregory and Matthew. Your support made much of this work possible and I am grateful to have had the opportunity to work with you these last 3 years.

Finally, I would like to gratefully acknowledge Canterbury Health Laboratories and the Maia Health Foundation for their financial support throughout the duration of this work.

Publications related to this thesis

Kennedy H, Haack TB, Hartil Vl, Matakovi L, Baumgartner ER, Potter H, Mackay R, CL Alston CL, OSullivan S, McFarland R, Connolly G, Gannon C, King R, Mead S, I Crozier I, Chan W, Florkowski CM, Sage M, Hfken T, Alhaddad B, Kremer LS, Kopajtich R, Feichtinger RG, Sperl W, Rodenburg RJ, Minet JC, Dobbie A, Strom TM, Meitinger T, George PM, Johnson CA, Taylor RW, Prokisch H, Doudney K, Mayr JA. 2016 Sudden Cardiac Death due to Deficiency of the Mitochondrial Inorganic Pyrophosphatase PPA2. (2016) AJHG.

Stevens AJ, **Kennedy HL**, Kennedy MA. Fluorescent methods for probing G-quadruplex structure in single and double stranded DNA. Biochemistry. 2016 Jun 2.

Aung NN, **Kennedy H**, Faed JM, Brennan SO. Novel heterozygous Bbeta (c.1311T>A) mutation (Fibrinogen St Kilda) associated with recurrent pregnancy loss. Pathology-Journal of the RCPA. 2015 Oct 1;47(6):583-5.

Kularatnam GA, Warawita D, Jayasena S, Nadarajah S, Jasinge E, Mendis D, **Kennedy H**, Florkowski C, George P. Cystic fibrosis in a Sri Lankan infant, confirmed by genotyping: implications for future diagnosis and service provision. Journal of the Postgraduate Institute of Medicine. 2015 Jul 27;2.

Manuscripts in preparation

Kennedy H, Doudney K, George P. *De novo* mutation in BCOR identified in a patient with an FA-like syndrome.

Contents

Attribution of collaborative contributions **i**

Abstract **iii**

Acknowledgements **v**

Publications related to this thesis **vii**

List of Abbreviations **xvii**

1 Introduction **1**
- 1.1 Overview . . . 1
- 1.2 History of molecular diagnostics . . . 1
  - 1.2.1 Early methodology . . . 2
- 1.3 DNA Sequencing . . . 3
- 1.4 Human Genome Project . . . 4
- 1.5 Emerging Non-Sequencing Technologies . . . 5
  - 1.5.1 SNP arrays . . . 5
  - 1.5.2 Genome Wide Association Studies (GWAs) . . . 6
  - 1.5.3 ArrayCGH . . . 6
- 1.6 Massively Parallel Sequencing (MPS) . . . 7
  - 1.6.1 MPS sequencing approaches and platform development . . . 7
  - 1.6.2 WES and WGS . . . 12
- 1.7 MPS Data analysis . . . 14
  - 1.7.1 General steps and available software . . . 14
- 1.8 Challenges . . . 18
- 1.9 Implementation . . . 19
- 1.10 Sequencing approaches . . . 20
  - 1.10.1 Disease-targeted gene panels . . . 20
  - 1.10.2 Exome sequencing . . . 21
- 1.11 Platforms . . . 22
  - 1.11.1 Illumina . . . 22
  - 1.11.2 IonTorrent . . . 23
  - 1.11.3 Automation of template preparation . . . 24
  - 1.11.4 Third party - library prep, automation and analysis . . . 25
- 1.12 Test development . . . 25
- 1.13 Validation . . . 27
- 1.14 Quality Control and Proficiency Testing . . . 28

1.15 Technical considerations . . . 28
1.16 Data analysis . . . 29
1.16.1 Incidental Findings . . . 30
1.16.2 Finding of uncertain significance . . . 30
1.16.3 Computational predictive programs . . . 31
1.17 Reporting . . . 32
1.18 Thesis direction . . . 33

2 Materials and Methods 35
2.1 Suppliers and nomenclature . . . 35
2.1.1 Equipment, reagents and consumables . . . 35
2.1.2 Nomenclature . . . 35
2.2 Clinical Samples . . . 35
2.3 General Methods . . . 36
2.3.1 DNA extraction . . . 36
2.3.2 DNA quantification . . . 37
2.3.3 Standard Polymerase Chain Reaction . . . 38
2.3.4 Agarose gel electrophoresis . . . 38
2.3.5 CGH Array . . . 39
2.3.6 Automated DNA sequencing methods (Sanger) . . . 39
2.3.7 Sequencing reaction purification . . . 39
2.3.8 Sequencing analysis . . . 40
2.4 Methods used in Chapter 3
Analysis of an intellectual disability pedigree . . . 40
2.4.1 Ion Proton sequencing methods . . . 40
2.4.2 Ion ProtonTM Sequencing . . . 43
2.4.3 Variant Analysis . . . 43
2.4.4 Resequencing . . . 43
2.5 Methods used in Chapter 4
Trio analysis - a sporadic case of premature aging . . . 45
2.5.1 Variant Analysis . . . 45
2.5.2 Resequencing . . . 46
2.5.3 Allele specific amplification of BCOR . . . 46
2.5.4 X-inactivation analysis . . . 47
2.6 Methods used in Chapter 5
A case of sudden cardiac death . . . 49
2.6.1 MiSeq sequencing methods . . . 49
2.6.2 Sequencing protocol . . . 49
2.6.3 Variant annotation . . . 49
2.6.4 Variant filtering . . . 49
2.6.5 Resequencing . . . 50
2.6.6 Structural modelling . . . 50
2.7 Methods used in Chapter 6
Wider investigations of PPA2 . . . 51
2.7.1 Structural modelling . . . 51
2.7.2 Oxidative stress sensitivity of PPA2 knockout yeast . . . 51
2.8 Methods used in Chapter 7
Zebrafish modellling of PPA2 dysfunction and mitochondrial disease . . . 52

2.8.1 Zebrafish husbandry 52
2.8.2 Preparation of ppa2 RNA for microinjection 53
2.8.3 Preparation of PPA2 RNA for microinjection 53
2.8.4 Morpholino preparation 53
2.8.5 Microinjection of mopholino and/or RNA 54
2.8.6 Dechorionation of embryos 54
2.8.7 Characterisation of phenotype 54

3 Analysis of an intellectual disability pedigree 57
3.1 Introduction 57
3.1.1 Case history 58
3.1.2 Proband exome analysis 59
3.1.3 *DACH2* Analysis 60
3.2 Multi-sample exome analysis 61
3.2.1 *KMT2C* Analysis 63
3.3 Discussion 67

4 Trio analysis - a sporadic case of premature ageing 73
4.1 Introduction 73
4.2 Case history 74
4.3 Exome analysis 75
4.3.1 Recessive analysis 75
4.3.2 *De novo* analysis 75
4.4 Investigation of *BCOR*, BCL6 co-repressor 77
4.4.1 Sanger confirmation 77
4.4.2 Allele-specific Sanger sequencing 77
4.4.3 Frequency and mutation effect 79
4.4.4 *BCOR* functional role 82
4.4.5 *BCOR* disease association 83
4.5 X-inactivation 85
4.6 FA exclusion 87
4.7 Discussion 90

5 A case of sudden cardiac death 99
5.1 Family Christchurch 99
5.1.1 Case history 99
5.1.2 Array CGH 102
5.1.3 Targeted genetic analysis 102
5.1.4 DNA preparation 103
5.1.5 Exome analysis 103
5.1.6 Sanger sequencing 106
5.2 *PPA2* variant investigations 106
5.2.1 *In silico* predictions of *PPA2* variants 106
5.2.2 Modelling 107
5.3 *In silico* prediction of mitochondrial variant 107
5.4 Discussion 111

6 *PPA2* and mitochondrial disease 119
6.1 Intro . . . 119
6.2 Case details . . . 119
 6.2.1 Family N Ireland . . . 124
6.3 PPA2 activity . . . 125
 6.3.1 Western blotting of PPA2 protein in human tissue . . . 125
 6.3.2 *ppa2* knockout growth phenotype on aerobic media . . . 126
 6.3.3 Oxidative stress sensitivity . . . 126
 6.3.4 Pyrophosphatase activity of recombinant PPA2 . . . 127
6.4 Discussion . . . 128
 6.4.1 *PPA2* in a mitochondrial context . . . 128
 6.4.2 Theoretical proposal of *PPA2* disease mechanism . . . 131
 6.4.3 Alcohol sensitivity and *PPA2* . . . 135
 6.4.4 Proposed Mechanism for SCD due to *PPA2* mutations . . . 139
 6.4.5 Conclusion . . . 142

7 Zebrafish modelling of PPA2 dysfunction and mitochondrial disease 143
7.1 Introduction . . . 143
7.2 Zebrafish *ppa2* . . . 145
7.3 Morpholino knockdown - preliminary experiments . . . 145
7.4 RNA Rescue with wildtype *ppa2* . . . 147
7.5 Additional phenotype characterisation methods . . . 151
 7.5.1 Heart related features . . . 152
7.6 CRISPR/cas9 targeted genome editing . . . 153
7.7 Discussion . . . 154

8 Conclusions and future directions 157
8.1 Analysis of an intellectual disability pedigree . . . 157
8.2 Trio analysis - a sporadic case of premature aging . . . 157
8.3 Analysis of a pedigree with sudden cardiac death . . . 158
8.4 Recommendations from thesis . . . 160
8.5 Summary . . . 162
8.6 Future directions . . . 163

Appendices 165

Appendix A Primer Sequences 167

Appendix B Zebrafish experimental details 169

Appendix C Gene lists used in exome analysis 173

List of Figures

1.1 Principle of Massively Parallel Sequencing 8

3.1 Pedigree with multiple affected individuals with ID 58
3.2 Amplification and segregation of *DACH2*:p.Glu415Asp 61
3.3 Mechanism of Recombination 62
3.4 Sibling exome variant filtering - Autism and ID panel script 64
3.5 Sibling exome variant filtering - IonReporter 64
3.6 Sequencing of *KMT2C*; p.Leu291Phe 66
3.7 Sequencing of *KMT2C*; p.Gly838Ser 66
3.8 Chromosome 21 homologous region (BAGE) 67
3.9 *KMT2C*; p.Gly838Ser allele specific primer design 68
3.10 *KMT2C*; p.Leu291Phe allele specific primer design 69
3.11 KMT2C- specific Sanger sequence 70

4.1 Pedigree and morphology 74
4.2 *BCOR*; p.Gly1303Ser - IGV visualisation 77
4.3 *BCOR* (Gly1303Ser) confirmation 78
4.4 *BCOR* allele-specific nested PCR optimisation 80
4.5 *BCOR* nested PCR-template dilutions 81
4.6 *BCOR* allele specific PCR optimisation 81
4.7 *BCOR* associations 82
4.8 The BCL6 complex 83
4.9 *BCOR* expression in tissues 84
4.10 *BCOR* protein annotation 86
4.11 X-inactivation assay for non-random skewing 88
4.12 X-inactivation assay of skin biopsy 89
4.13 Establishing *de novo* mutation causality 92

5.1 Family Christchurch pedigree showing cardiac condition 100
5.2 Cardiac MRI 101
5.3 Heart section 101
5.4 Variant filtration 105
5.5 *PPA2* variants - IGV visualisation 106
5.6 Variant resequencing and modelling 108
5.7 Yeast structural model - Glu172 109
5.8 Yeast structural model - Pro228 110
5.9 COXIII structural conservation 111
5.10 Dimer combinations 113

5.11 OXPHOS schematic 114

6.1 *PPA2* pedigrees 120
6.2 Human PPA2 and Yeast PPA2 protein alignment 126
6.3 Western blot analysis and yeast *PPA2* knockout strains 127
6.4 PPA2 enzyme activity in fibroblast mitochondria 129
6.5 Proposed pathogenic mechanisms of PPA2 dysfunction 136
6.6 Proposed influence of ethanol ingestion on PPA2 dysfunction 137
6.7 *PPA2* mechanism of disease 141

7.1 Human PPA2 and Zebrafish PPA2 protein alignment 145
7.2 Morpholino sequence details 146
7.3 Morpholino 148
7.4 Morpholino 149
7.5 Morpholino 150
7.6 Morpholino 151
7.7 Zebrafish ventrical shortening fraction 152
7.8 Zebrafish heartrates 153
7.9 CRISPR/cas9 fry 154

B.1 pExpress-1 170
B.2 pCS2+ Plasmid 171
B.3 pCS2+ cloning plan 171

List of Tables

1.1 Commercial gene panels . . . 26
1.2 Custom gene panels . . . 26
1.3 Thesis aims . . . 34

2.1 Proband-only X chromosome variant filtering . . . 44
2.2 Sibling variant filtering . . . 44
2.3 Recessive analysis filtering pipeline . . . 46
2.4 *De novo* filtering pipeline . . . 47

3.1 Candidate variant details . . . 59
3.2 *In silico* pathogenicity predictions for *DACH2*:p.Glu415Asp . . . 60
3.3 *In silico* prediction of pathogenicity for *KMT2C* . . . 65

4.1 Candidate variant - recessive . . . 75
4.2 Candidate variants - sporadic . . . 76
4.3 *In silico* prediction of pathogenicity for *BCOR*; p.Gly1303Ser . . . 79

5.1 Cardiomyopathy-associated gene list . . . 104
5.2 *In silico* predictions . . . 107

6.1 Cellular reactions requiring PPA2 involvement . . . 134

A.1 Primers used in general methods . . . 168
A.2 Primers used in Chapter 3 . . . 168
A.3 Primers used in Chapter 4 . . . 168
A.4 Primers used in Chapters 5-6 . . . 168

B.1 Morpholino sequences . . . 170

C.1 Intellectual disability (ID) associated genes . . . 173
C.2 Autism spectrum disorder (ASD) associated genes . . . 174

List of Abbreviations

ACMG	American College of Medical Genetics and Genomics
AML	Acute myeloid leukemia
ATP	Adenosine triphosphate
AU	Absorbance unit
BDT	BigDye terminator
bp	Base pair
BPB	Bromophenol blue
°C	Degrees celsius
CGH	Comparitve genomic hybridisation
CNV	Copy number variant
dfp	Days post fertilisation
DMSO	Dimethyl sulfoxide
EBV	Epstein-Barr virus
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encylopedia of DNA Elements
ETC	Electron transport chain
ExAC	Exome Aggregation Consortium
FA	Fanconi anaemia
FFPE	Formalin-Fixed Paraffin-Embedded
g	Gram
Gb	Gigabase
gDNA	Genomic DNA
GUS	Gene of uncertain significance
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hpf	Hours post fertilisation
ID	Intellectual disability
IF	Incidental finding
IGV	Integrated Genomics Viewer
Kb	Kilobase
LE	low electroendosmosis
Mb	Megabase

MMPT	Mitochondrial membrane permiability transition
MO	Morpholino
MPS	Massively parallel sequencing
MPTP	Mitochondrial permiability transition pore
MRI	Magnetic ressonance imaging
mRNA	messenger RNA
mtDNA	Mitochondrial DNA
NAD	Nicotinamide adenine dinucleotide
nDNA	nuclear DNA
ng	nano gram
NGS	Next generation sequencing
NTP	nucleoside triphosphates
OD	Optical density
OFCD	Oculofaciocardiodental
OXPHOS	Oxidative phosphorylation
PCR	Polymerase Chain Reaction
pg	Pico gram
Pi	Inorganic phosphate
PolyPhen	Polymorphism Phenotyping
PPi	Inorganic pyrophosphate
QC	Quality control
ROS	Reactive oxygen species
rpm	Revolutions per minute
SCD	Sudden cardiac death
SIFT	Sorting Intolerant from Tolerant
SNP	Single nucleotide polymorphism
TAE	Tris-acetic acid EDTA
TBE	Tris-borate EDTA
TCA	Tricarboxylic acid cycle
TE	Tris-EDTA
Tris-HCL	Tris hydrochloride
tRNA	transfer RNA
uL	micro litre
uM	micro molar
UTR	Untranslated region
VD	ventricular diastole
VS	ventricular systole
VUS	Variant of uncertain significance
WAIS	Wechsler adult intelligence scale
WASP	Web-based allele specific primer
WES	Whole exome sequencing
WGS	Whole genome sequencing
WT	Wildtype
x g	Times gravity
XC	Xylene cyanol

Chapter 1

Introduction

1.1 Overview

Genomics is a relatively new term that encompasses the study of genes, their functions and related techniques [2]. This chapter discusses the journey of molecular diagnostics from single gene and single mutation diseases, to the rapidly evolving world of Massively Parallel Sequencing (MPS) and addresses the main issues that need to be considered when transferring this research technique into a routine diagnostic tool.

1.2 History of molecular diagnostics

Genetic disease places a considerable health and financial burden on patients and affected families, as well as the wider community. Genetic disorders have been estimated to account for thirty percent of admissions to paediatric wards and many lead to paediatric mortality [3]. Of those patients that survive into adulthood, many are living with severe and incurable disease. The emphasis is often placed on prevention, of both complications related to the disease in the individual, and recurrence within the family through genetic counselling. For both of these cases, an accurate diagnosis is crucial [4]. Online Mendelian Inheritance in Man (OMIM), an authoritative knowledge base of human genetics and Orphanet (a catalogue for rare disease), predicts that there are between 6,000 and 7,000 rare monogenic disorders, however, only half of these currently have a described molecular aetiology [4, 5].

1.2.1 Early methodology

Molecular disease was a phrase first coined by Linus Pauling based on discoveries relating to sickle cell disease (SCD). In 1949 Pualing *et al* [6] demonstrated that SCD was caused by a defect in the haemoglobin molecule and inheritance of the disease was shown to be autosomal recessive. In 1958 the single amino acid substitution in the beta globin chain that leads to this abnormal haemoglobin was characterised [7]. These early experiments put SCD at the leading edge of molecular studies in human disease. SCD falls into a class of disorders called haemoglobinopathies, which also includes thalassemias. Haemoglobinopathy implies structrual abnomalities in the globin protiens, whereas thalassemias usually result in underproduction of of normal globin proteins [8]. The most clinically relevant thalassemias are the α and *beta* thallassemia [8]. Mutations in the globin genes can either lead to both thalassemias characterised by decreased production of normal globin chains, and variant haemoglobin molecules carrying amino acid substitutions. Most mutations that give rise to haemoglobinopathies are either point mutations or small deletions, and only rarely insertions. β-thalassemia is mostly caused by point mutations (some located within deep intronic regions), while α-thalassemia is mostly caused by large deletions. The deletions associated with α-thalassemia, and Hereditary Persistence of Fetal Haemoglobin (HPFP) were first characterised using the technique of Southern blotting [9]. The advent of recombinant DNA technology, developed in the 1970s, made it possible to characterise DNA molecules directly and provided the basic knowledge on the primary sequence of many genes [10]. Point mutations that cause β-thallassemia were elucidated in this way [11]. Molecular cloning techniques, and recombinant DNA technology vastly accelerated the rate at which genes from complex organisms could be studied. Pioneered by microbiologists, this technology arose from the use of restriction endonucleases to cleave DNA at specific DNA sequences, and purification of specific fragments by size fractionation [12]. DNA ligase enabled these fragments to be assembled in new combinations. Recombining DNA into vectors such as plasmids or bacteriophage resulted in the ability to naturally replicate the DNA fragment in a bacterial culture. Initially proposed by Peter Lobban of the Stanford University Medical School, the first papers describing successful recombinant DNA experiments were published in 1972 and 1973 [13, 14, 15, 16]. By 1976, prenatal diagnosis of α-thalassemia, using hybridisation of fetal fibroblast DNA had been successfully performed [17]. Throughout the 1980s, similar diagnostic approaches allowed prenatal testing of several other genetic disorders including cystic fibrosis [18] and phenylketonurea [19].

Ten years after the first successful experiments, recombinant DNA technology was used to generate human insulin, the first instance of a recombinant licensed drug [20].

Despite these rapid advances, the identification of disease causing mutations in this era was laborious and time consuming. Routine molecular diagnosis of inherited disease was not a feasible option in clinical laboratories until the development of the polymerase chain reaction (PCR) technique in the late 1980's [21, 22, 23]. This technology revolutionised molecular diagnostics, allowing easy DNA sequence analysis of selected 'candidate' regions of specific genes and identification of known mutations within a single day.

1.3 DNA Sequencing

DNA sequencing was introduced in 1977, in two seminal papers by Allan Maxam and Walter Gilbert [24], and Frederick Sanger and colleagues [25]. The Maxam-Gilbert method involved base-specific chemical cleavage of radioactively labelled DNA fragments, and separation by gel electrophoresis prior to visualization using autoradiography. The Sanger method incorporated chain-terminating dideoxynucleotides in the sequencing reaction which caused base specific termination of primed DNA synthesis. The Sanger method prevailed as the more robust technique, and became routinely adopted. Further development replaced radioactivity with fluorescent labelled primers, then later on with fluorescent labelled dideoxynucleotide chain terminators [26, 27]. In addition, fragment separation was converted from electrophoresis on a polyacrylamide gel followed by transfer onto filter paper and drying, to capillary electrophoresis (CE) and laser detection of the nucleotide fluorescent signal. This allowed the compression of four lanes previously required on a polyacrylamide gel into one capillary, and analysis of multiple samples simultaneously with a capillary array. This enabled rapid commercialisation of the

Sanger method and a leap in the sequencing production capabilities of laboratories.

1.4 Human Genome Project

The Human Genome Project (HGP) was an international, collaborative research effort, with the key goal to detect map and identify every human gene. Initially proposed in 1985, the merit of the project's concept was fiercely debated [28]. The HGP was originally viewed as discovery science, rather than hypothesis driven science, where the experiment(s) validated a pre-formed hypothesis. One popular belief was that the non-gene-coding parts of the genome (98%) were largely irrelevant 'spacing' DNA, and that the project would provide limited useful data. Additionally, the technological components for high-throughput automated DNA sequencing and handling of the data were not available and would need to be developed. In 1998 the National Academy of Science appointed a committee to consider the proposal.

After a year of debate and analysis, the committee unanimously endorsed the project.

The project was given a 15-year time frame, with an initial budget of $3 million to achieve six distinct objectives:

1. To generate physical, genetic, and sequence maps of the human genome.
2. To sequence the genomes of a variety of model organisms.
3. To develop and improve the technologies for mapping and sequencing.
4. To develop computational tools for capturing, storing, analyzing, displaying, and distributing the data.
5. To sequence expressed sequence tags (ESTs) and cDNAs to establish expression of mRNAs in different cell types of both humans and mice.
6. To consider the ethical, legal, and social implications (ELSI) posed by genomic information.

The government-funded project was initially driven by the International Human Genome Sequencing Consortium (IHGSC), and in 1998 a private company Celera joined the race with a novel sequencing strategy and the promise of a draft in just three years. The competition prompted a revision of the IHGSC time frame, and eventually resulted in the project finishing two years ahead of schedule. The IHGSC published the first draft of the human genome in February 2001 [29], the same month as Celera Genomics [30]. A completed description of the finished human genome sequence by the IHGSC was published in Nature in October 2004 [31]. This paper provided the scientific community with an assessment of the quality of sequence that had been deposited in public databases in April of the previous year.

Up until 2001 the human genome had been sequenced on instruments running capillary electrophoresis of individual fluorescent labeled Sanger sequencing reaction products. These early instruments were limited to analysis of 500-600 base reads, from 96 reactions in approximately 10 hours. The first five years post publication of the draft genome also saw the addition of drafts or finished sequences for several model organism genomes. These included *Saccharomyces cerevisiase* [32], *Escherichia coli* [33],*Drosophila melanogaster* [34], and *Caenorhabditis elegans* [35], and then, shortly after; genomes of the mouse [36], rat [37], domestic dog [38], chicken [39], chimpanzee [40],rhesus macaquet [41], and cow [42]. These genomes provided further annotation of the human genome by comparative genomics; assuming highly conserved genomic elements between species were more likely candidates for functional importance [43].

1.5 Emerging Non-Sequencing Technologies

Completion of the HGP provided a foundation for modern biomedical research and clinical molecular diagnostics. Although advances in sequencing were perhaps most crucial to the development of genomics, there are several other genome scale techniques that were developed, including SNP arrays, Genome-Wide Association Studies (GWAS) and ArrayCGH (comparative genomic hybridisation).

1.5.1 SNP arrays

SNP arrays were originally designed to genotype human DNA at thousands of polymorphic loci across the genome simultaneously. The Affymetrix SNP array was the first to be commercially produced, and was able to genotype 1494 SNPs on one chip. Current SNP arrays have greatly improved on this, and contain nearly 1 milllion SNPs [44].

1.5.2 Genome Wide Association Studies (GWAs)

Genome wide association studies that use SNP arrays to detect millions of variants in a single DNA sample became a popular method for detecting association between SNPs and a disease or trait. This technology requires comparison of a cohort of healthy controls (that do not show the trait or disease of interest) to a cohort of affected patients. Any alleles (or group of alleles) that are significantly more frequent in the affected group are described as associated, and indicate that the region of DNA in which the SNP is located may influence the risk of disease. This is based on the principle of linkage disequilibrium (the non-random association of alleles at physically separate loci) at the population level. The first GWAS was published in 2005 [45] which investigated alleles associated with age related macular degeneration. The Wellcome Trust Case Control Consortium (WTCCC) study in 2007 [46] was another landmark project, and was the first GWAS designed for complex disease that used a chip with good coverage of the whole genome. This study included 14000 cases, covering 7 common diseases; coronary heart disease, rheumatoid arthritis, Chrone's disease, type 1 diabetes, type 2 diabetes, hypertension and bipolar disorder. The number of individuals included in these studies has increased dramatically since these early studies. An example of a more recent sample size is 200 000 individuals [47] investigating blood pressure and cardiovascular disease risk.

1.5.3 ArrayCGH

ArrayCGH (also referred to as microarray-based CGH) is a technique for detecting chromosomal copy number variants on a genome-wide scale. The methodology is based on comparison of a patient's genome to that of an apparently normal reference genome, which allows the detection of regions of imbalance as with gains or losses of copy number. This technology extended the resolution of the traditional cytogenetic G-banding CGH technique which could only detect larger chromosomal abnormalities in metaphase chromosomes (in the region of >10Mb). In arrayCGH, metaphase chromosomes are replaced by specific DNA fragments which act as probes covering the entire non-repetitive genome. This enables identification of microdeletions/duplications at a resolution of 10-50kb. This has applications in cancer genetics [48], and for a multitude of syndromes caused by submicroscopic aberrations such as Prader-Willi/Angelman syndrome [49], CHARGE syndrome [50], DiGeorge syndrome [51] and Williams-Beuren syndrome [52], among many others.

1.6 Massively Parallel Sequencing (MPS)

Although the Human Genome project was a huge accomplishment, the 13-year effort cost an estimated $2.7 billion [29, 31]. Within five years of completion, new sequencing technologies were able to replicate this feat for approximately $1.5 million over a five month period [53]. These new technologies were termed 'next-generation' sequencing (NGS) or 'massively parallel' sequencing (MPS) and have revolutionised genomic research, enabling experiments that were previously not technically feasible or affordable [54]. The principle behind MPS is a paradigm shift from that of Sanger sequencing. Instead of performing individual capillary electrophoresis reactions, MPS employs array-based sequencing which can process millions of reactions in parallel. This method can generate gigabases of nucleotide-sequence output in a single instrument run. The general process of MPS is illustrated in Figure 1.1 although the specific methodology differs between the platforms and chemistries which are now commercially available.

1.6.1 MPS sequencing approaches and platform development

Reversible dye terminators

One of the initial sequencing chemistry approaches used in massively parallel sequencing were reversible dye terminators.

Solexa developed their system in 2007, although it was quickly acquired by Illumina, Inc [55]. The library preparation includes fragmentation of high-molecular weight DNA, enzymatic trimming, and adenylation of the fragment ends and ligation of specific adapters. Bridge amplification of the library is performed on a microfluidic conduit (a flow cell) made of flat glass with 8 channels, coated with covalently bonded adapter sequences complementing the library adapter sequences. These fragment ends are released post amplification to allow priming with comple-

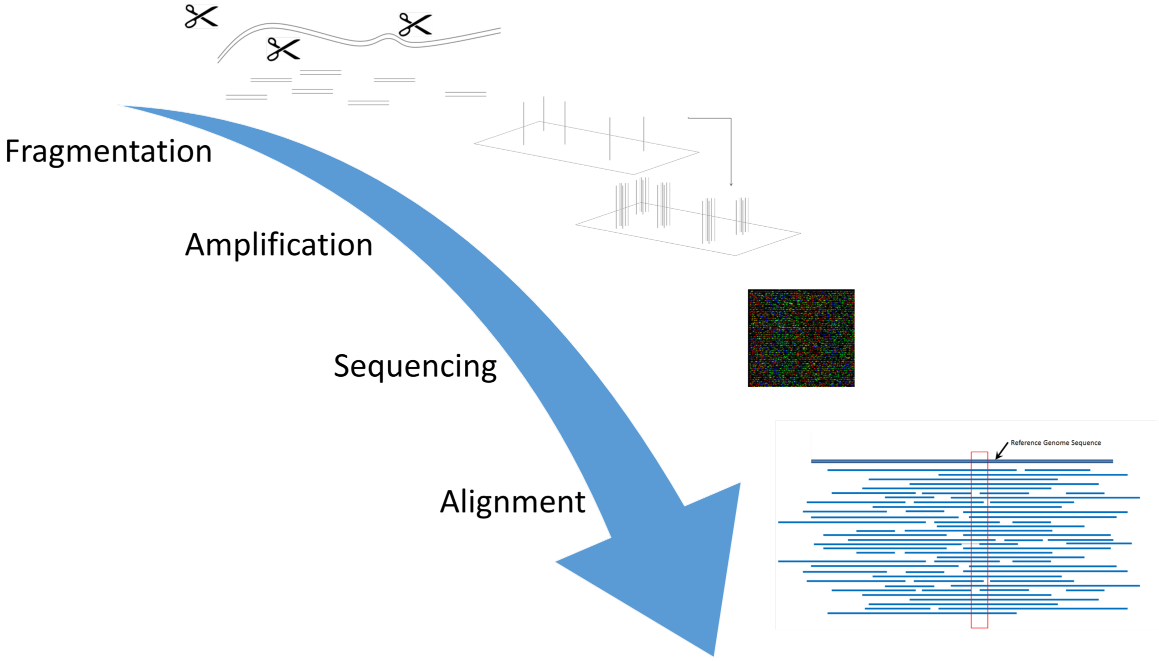


Figure 1.1 General steps involved in MPS
A Fragmentation; Random enzymatic fragmentation of the input genomic material into short segments is required before ligation or chemical binding of non-specific adapter DNA sequences occurs. These ligated DNA fragments are either bound to a solid surface, or are captured by micron-scale beads. **B** Amplification; Depending on the adapter ligation approach, PCR is either carried out in emulsion, or on a surface in a bridging reaction. This step amplifies the DNA fragments in a non-specific clonal fashion, amplifying the signal that will be detected in subsequent sequencing. **C** Sequencing; Techniques for base calling are platform specific, but generally rely on either a fluorescent signal or the pH change that can be measured when a base is incorporated into a growing chain. **D** Alignment; Overlapping sequencing reads are aligned and mapped back to a reference sequence computationally.

mentary synthetic DNA that can be extended in subsequent sequencing reactions. The use of reversible dye terminators dictates that the addition of nucleotides is performed in a step-wise fashion. All four uniquely fluorescent labelled nucleotides are provided in each cycle, however a blocking group at the 3-OH position of the ribose sugar prevents additional base incorporation after one has been added. After this cycle, unincorporated nucleotides are washed away and the flow cell is imaged to identify clusters that are reporting a fluorescent signal. The florescent groups are then chemically cleaved and the 3-OH blocker removed. This sequence is repeated for up to 150 nucleotide reactions, at which point the reads are resequenced from the opposite end of each fragment cluster [56].

Ion detection

An alternate to fluorescence was initially produced by Ion Torrent (who were later purchased by Life Technologies, then Thermo Fisher Scientific). This technique utilises measurement of the pH change caused by the release of hydrogen ions during nucleotide incorporation. This removes the requirement of florescence chemistry, and instead relies on a novel silicon detector. Library preparation for this technique involves emulsion PCR. Here, fragments of a genomic DNA library are emulsified with small beads (with covalently linked adapter complementary sequences on their surface), PCR reactants, and DNA polymerase molecules. The emulsion produces hundreds of thousands of individual oil micelles, each containing one bead and allowing compartmentalised PCR amplifications within one mixture. Post amplification, the beads must be released from the emulsion, and the library enriched for beads that were successfully amplified. Additional priming steps involve annealing a sequencing primer, and loading of the beads into the wells of a Ion Chip. This chip is specifically designed to act as the microfluidic conduit, and to detect the pH changes occurring in each well. In this system the nucleotides are flowed across the chip sequentially, as there is no florescent label to provide base specificity [56].

Zero-mode waveguide

In 2010 Pacific Biosciences commercialised a single-molecule sequencer that uses the zero-mode waveguide (ZMW), a nanotechnology structure small enough to allow tens of thousands to be manufactured on a silicon wafer array [57]. The ZMW focuses light, and when bound to DNA polymerase on a surface the instrument excitation/detection optics can be trained at the attachment site. When fluorescent nucleotides are added to the surface, the instrument detects the presence of each nucleotide that enters the active site in real time, as the polymerase is copying the template. This allows read lengths of up to 10,000 nucleotides at a time, although as yet the error rate is much higher (15%) than the amplification based approaches [56]. It should be noted that these errors are more likely to be random, as apposed to the systematic errors that can occur with enzymatic chemistries. Repetitive sequencing of a molecule will reduce the effect of random errors in the consensus sequence.

Third generation sequencing

Sequencing developments beyond the MPS methods described above have been coined as the 'Third generation' of sequencing. Currently these methods are typically in developmental stages and are not ready for diagnostic clinical applications. Third generation sequencing currently involves reading un-amplified single stranded nucleic acids as they pass through nanopores. This technology is based on the principle first demonstrated in 1996 [58], that when a voltage is applied across an ionic system, separated by an impermeable membrane that contains a nanopore, then ions are forced through the pore producing a current. When a charged molecule such as DNA enters the pore, the flow of ions is resisted and this creates a change in the electrical current that can be measured. If the 4 bases (A,T,G,C) are all of differing sizes and would resist the ion flow uniquely, then each base would produce a electrical signature that would allow sequencing of long strands of template. There are two major companies taking advantage of biological nanopores: Oxford Nanopore Technologies (ONT), and Genia, and commercialisation of their single molecule sequencing instruments is well under way. Both an MspA (Mycobacterium smegmatis porin A) pore [59], and an α-hemolysin pore [60] have been shown to be effective in DNA sequencing proof of concept papers [61].

Biological nanopores have the advantage of being easily modifiable, and able to be produced in mass with consistent size and structure. In the case of the α-hemolysin pore, genetic modifications were required before use in sequencing. These involved the addition of a specific mutation within the beta barrel to achieve base discrimination, and non-covalent attachment of a cyclodextrin (CD) derivative which slowed the translational of the DNA bases through the nanopore sufficiently to allow accurate detection [62]. In the MpsA pore, double stranded DNA is used to block translational in place of CD. This requires extensive pre-sequencing preparation of strands in order to separate blocks of target ssDNA with dsDNA. A development on this method is to use DNA polymerase (DNAP) phi29, which acts as a ratchet, pulling single stranded template through the pore one base at a time [63]. This DNAP has been shown to work in the α-hemolysin pore too, and this work has been licensed by ONT [64]. The ultimate aim is to accelerate new and revolutionary DNA sequencing methods and technologies. ONT are partners of READNA and their first platforms (the GridION, and the disposable version the MinION) were commercialised to a select number of customers in late 2012. Early publications of the MinION suggested that the device had too high an error rate to be of utility in genotyping applications [65], however more recent studies have shown an increased accuracy with advances in sequencing chemistries [66]. The major advantage of nanopore sequencing is evident in the length of align-able sequence reads, reaching >100kb [67] (vs 500bp reads seen with Illumina and IonTorrent chemistries) and it is expected that sequencing accuracy and data output will continue to improve. More recently the concept of DNA origami has been introduced, as a method of producing nanopores with greater chemical and biological specificity, and near atomic precision [68]. The other path being pursued is the use of solid state nanopores, although biological nanopores have been more extensively studied.

Forth generation sequencing

The so called 4th generation of DNA sequencing focusses on transcript analysis of cells in parallel in their cellular context (*in situ* sequencing). This sort of development is predicted to become important especially in tumour sequencing and is already on the distant horizon [69].

1.6.2 WES and WGS

Whole genome sequencing (WGS) and whole exome sequencing (WES) are both unbiased and powerful MPS approaches for detecting genetic variation [70]. WGS covers the entire sequenceable genome, intronic and exonic. WES is amplification and sequencing of the exonic portion of the genome, achieving greater depth at any given base pair for the same amount of genetic data. Although only a small proportion of the genome is interrogated by this method (approximately 1%), exonic variants are significantly more likely to cause disease than those identified in the remaining 99% of the sequence. Exome sequencing is of particular use in cases of rare, Mendelian disease, as focusing on the coding regions often yields relevant variants at a fraction of the cost of WGS. At the time of writing this thesis, however, the balance is shifting towards WGS becoming more cost effective and actually providing more depth of cover than WES.

Oligocapture vs PCR Amplification

In order to examine specific regions of the genome, several 'target-enrichment' methods have been developed which selectively isolate regions of interest prior to sequencing. Although not exclusive, there are two methods that are most commonly employed, PCR amplification and oligocapture.

PCR amplification has been a key sample preparation technique for over 20 years, and has been applied with success to generating MPS libraries. For the application of WES, amplification by PCR of only the exonic regions quickly enriches the target. This is often performed in highly multiplexed reactions, or emulsion PCR which allows single tube amplification of large numbers of independent targets. One drawback of PCR target enrichment methods is the difficulties in designing primers that will amplify 100% of the target region and a relatively high input requirement of genomic material. Additionally PCR enrichment will result in variable depth and systematic errors in the sequence. In contrast to amplification of a target region, oligocapture target-enrichment involves hybridisation of a fragment library to specific probes that target regions of interest. This capture can occur on a solid surface or in solution. Both methods have their advantages and disadvantages, and often the sample size will dictate the method [71]. The fact that oligocapture can more accurately detect large deletion and duplication events (without PCR bias) may make this a more attractive option for diagnostic arrays. However, an amplification panel can more readily be updated to include new regions of interest, and currently this may be a more cost effective method.

Early application of exome sequencing

The utility of WES was first demonstrated in 2010 with the identification of genes responsible for Freeman-Sheldon syndrome (FSS)[72], Miller syndrome [73], and Schinzel-Giedion syndrome [74]. FSS has several characteristic features including a small pursed mouth, midface hypoplasia, and multiple contractures in the hands and feet. In this study, four unrelated affected individuals underwent WES and a list of variant-containing genes was generated. When variants were compared across data sets, only one affected gene was common to all four individuals; MYH3 [72].

This gene had been previously shown by linkage analysis to be causal for FSS, showing that massively parallel sequencing could be used to find causal mutations for monogenic disease even without linkage information, or any information relating to disease mechanism [75].

The first application of WES in a rare Mendelian disease of unknown cause was in Miller syndrome[73]. Affected individuals present with multiple malformations including severe micrognathia, cupped ears and missing 5th digits of the feet. This study identified DHODH as a candidate gene in four affected individuals from three independent kindreds. Further sequencing of three unrelated affected individuals showed that compound heterozygosity of mutations in DHODH was causative, and sequencing of parents indicated that this was a recessive disease not typically caused by sporadic mutations.

Following on from this success, WES was employed to determine the basis of Schinzel-Giedion syndrome, a distinctive disorder causing abnormal facial features, congenital malformations and severe mental retardation. It is usually fatal by 10 years of age. Exome sequencing of four unrelated affected individuals identified *de novo* heterozygous mutations in SETBP1 as causative in all four cases. Sanger sequencing of SETBP1 in additional patients with clinically diagnosed Schinzel-Giedion syndrome showed mutations in eight of the nine individuals investigated and confirmed the autosomal dominant mode of inheritance [74].

Autosomal recessive disorders are largely over-represented in early literature on massively parallel sequencing, with over 115 novel disease genes identified [76]. This is largely due to the relative simplicity of identifying rare, compound heterozygous mutations (in the absence of consanguinity) in only one affected sibpair [76].

1.7 MPS Data analysis

1.7.1 General steps and available software

As MPS rapidly generates a flood of heterogeneous data, management and analysis of the data has become crucial in avoiding the bottleneck in clinical utility. In order to efficiently perform the multiple analysis steps required, there are a large range of commercial and freeware programs, databases, and tools available.

For each different application of MPS one pipeline (or set of analysis steps) may be more suitable than another, and for very specific applications there may be particular workflows that are required. Choosing a pipeline or workflow to facilitate analysis of a particular application involves a comprehensive understanding of the stages of analysis and what choices are available.

Quality assessment

The first analysis step once data has been generated, is to evaluate the quality of the raw reads, and trim or remove reads that do not meet the required standard. This step typically addresses issues such as base calling errors, INDELs, adapter/primer contamination and poor quality sequencing at the ends of fragments [77].

There are multiple tools that can perform all or some stages of quality assessment, and many of these are inbuilt to the MPS platform being run. Comparison of these tools is outside the scope of this thesis, but the existance of so many methods suggests that all have their limitations or particular strengths. Of the platform-independent tools, NGSQC Toolkit [78], PRINSEQ[79], FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) and Galaxy [80] are some of the most commonly employed. NGSQC Toolkit and PRINSEQ are able to handle FASTQ and SFF input files, trim and filter, and produce a summary QC report. FastQC is compatible with all major MPS platforms and quickly produces summary tables and graphs to asses data quality. Galaxy is an integrated web-based tool that can create a FASTQ summary, as well as perform trimming and filtering tasks. SolexaQA47 and htSeqTools [81, 82] are two additional tools that include quality assessment, processing and visualisation functionality, htSeq Tools in particular is well suited to the high throughput requirements of ChIP-Seq or RNA-Seq studies while SolexaQA is designed for Illumina data.

Read alignment

Reads that pass a QC requirement are usually aligned to an existing reference genome. Both the University of California at Santa Cruz (UCSC) and the Genome Reference Consortium (GRC) are reliable sources for the human reference genome assembly. Both host multiple versions of the genome, UCSC currently provides hg18 and hg19, while GRC currently provides GRCh36 and GRCh37. The hg prefixed and GRCh prefixed references are identical at the nucleotide level but differ in nomenclature. The implementation of MPS required the rapid development of programs capable of aligning millions of short reads to a reference. The more commonly used tools to achieve this include BWA; [83], Bowtie/Bowtie2; [84], Novalign (http://novocraft.com), SOAP2 [85], MAQ [86], SSAHA [87] and YOABS [88]. Initial alignment programs were designed to deal with the short reads generated by early MPS runs. More recent software can deal with longer reads and gaps and largely follow the seed-and-extend method. This involves searching for one or more so-called seeds before extension to cover the whole query sequence [89].

Variant identification

One of the most critical steps in the clinical implementation of MPS is the variant identification stage. Tools for variant calling can be grouped into four applications: somatic, germline, copy number variant (CNV) and structural variant (SV) detection. Some key tools for germline mutation calling include GATK [90], CRISP [91], SAMtools [92] and SNVer [93]. GATK can be used as both a germline and somatic mutation caller. It is a software library that hosts several tools for working with MPS data including depth of coverage analysers, a local re-aligner, and SNP/INDEL caller. CRISP was developed to detect both rare and common variants from pooled MPS data. It is not intended for analysis of single samples, however. SAMtools is a collation of multiple tools that contains a subset of commands called BCFtools. These have the ability to call SNPs and short INDELs from a single alignment file, as well as identify somatic mutations. SNVer is primarily a statistical tool for the identification of SNPs and INDELs in both individual and pooled MPS data. These tools all differ slightly in which file formats they can manage, and which applications they are most useful for. Often, several callers may be combined in one analysis to ensure best coverage of potentially pathogenic mutations.

Annotation

The sheer size of the output files generated by MPS requires rapid and accurate automated variant annotation to filter and prioritise variants based on their functional impact and potential pathogenicity. The variants that can be annotated include SNPs, INDELs, and more recently CNVs, depending on the software and the application. The easiest way to provide annotation data is to link variants to entries in public variant databases such as dbSNP. There are various tools available for variant annotation, both web-based and off line. Some of the more common tools currently in use include ANNOVAR [94], SeattleSeq (http://snp.gs.washington.edu /SeattleSeqAnnotation138/), ExAC [95] and AnnTools (http://anntools.sourceforge.net/).

Visualisation

Visualisation of sequence properties is an important feature in any analysis pipeline. There are several browsers that enable loading of sequence data, and can display multiple tracks with user-specified information. Browsers enable visualisation of not only the aligned reads and annotated variants, but also quality information such as coverage and read depth. The browsers available generally fit into two categories; a web-based application, or stand-alone tool. The Ensembl Genome Browser [96] is a popular example of a web-based application that hosts multiple reference genomes, as well as local annotation sets and external resources. The University of California at Santa Cruz (UCSC) Genome browser [97] is also a web-based application that offers phenotype and disease annotations among its other uses. The UCSC database is comprehensive, containing nearly 1800 annotation tracts for the human genome GRCh37/hg19.

Stand-alone tools include Artemis [98], Intergrative Genomics Viewer (IGV) [99] and Sequence Annotation and Visualisation and Analysis tool (Savant) [100]. Artemis utilizes BamView [101] to display aligned reads. It allows filtering of variants based selected properties and the export of calculated properties such as SNP density and read counts [77] . IGV is a versatile tool, allowing import and the handling of >30 file formats. Reference genomes as well as data tracks and annotation from local or remote data sources can be viewed easily. The Savant tool allows loading of local as well as remote data sources and supports simple file formatting such as VCF zipping and indexing.

Prioritisation/Filtering

The process of identifying pathogenic mutations amongst the large number (typically 20,000) of variants generated by WES, can involve multiple approaches.

In general, for monogenic Mendelian disease the causative variant(s) will be rare in the population, likely to be located in a coding region, highly penetrant, of large effect and will segregate with disease. Initial filtering therefore would usually highlight missense and nonsense substitutions, as well as coding indels and splice site acceptor and donor sites. This list would then be further refined to those variants that are rare or novel. An additional method of filtering is based on the effect of a variant on protein structure and function, and how well this protein is conserved between species [75]. Several tools are available that assign pathogenicity scores based on one or more of these parameters. Sorting Intolerant From Tolerant (SIFT) [102], Polymorphism Phenotyping (PolyPhen) [103], PhyloP [104], and Genomic Evolutionary Rate Profiling (GERP) [105] are some examples. In many cases, potential causal variants will be identified, but further functional studies will often be required to show pathogenicity. All prioritisation approaches risk removing the pathogenic variant, which is reflected in the high rates of false-negative predictions (30% in one study of trios [106]) currently reported, and therefore the use of multiple prioritisation tools with phenotypic, pedigree and external (neutral and pathogenic) variant database information may be the most accurate approach. Removal of non-coding pathogenic variants will be a large contributor to false negative results.

1.8 Challenges

The current estimated number of heterozygous variants that will be found in any one sequenced genome, classified by the Human Gene Mutation Database as causing an inherited disorder is 50-100 [107]. This is in addition to any novel or uncharacterised variants that may be found in disease causing genes. This presents a challenge for not only correctly identifying the cause of disease, but also for counselling a individual to make them aware of the potential consequences of identifying a genetic variant that may be involved in future and/or unrelated disease. This has raised a number of ethical issues, such as the obligation of researchers to actively look for pathogenic variants or how individuals should receive their genome data back. In addition to known pathogenic variants, there will be the chance of finding variants of uncertain significance (VUS). As more human variation is discovered, there is a significant lag in accurate *in silico* prediction tools and functional studies to characterise these variants as benign, pathogenic or otherwise. Clear guidelines for distinguishing pathogenic sequence variants from the many benign changes in the human genome are urgently needed, particularly to reduce an acceleration of false-positive reports which would ”impede the translation of genomic research findings into the clinical diagnostic setting and hinder biological understanding of disease” [107]. The hugely increased amount of data now readily produced increases both the computing and analysis burden. Where previously a handful of genes were examined one by one, now thousands of genes are condensed into lists of variants (or variations from the reference sequence) which need to be filtered and prioritised. In the early stages of MPS these genomic changes may also need to be validated by a secondary method, usually Sanger sequencing, which can increase the workload for an individual patient considerably. The terabytes of data being produced puts an increased demand on resources for storage and handling, requiring more expensive in-house storage systems, or pushing data into the cloud which then causes issue regarding patient privacy.

On one hand there are many groups addressing these challenges to responsibly deal with genomic data, including the EuroGentest Validation Group [108], the CDC-convened Next-Generation Sequencing Standardization of Clinical Testing work group (Nex-StoCT) [109], the College of American Pathologist (CAP) [110] and many others [111, 109, 112, 113]. On the other hand there is the advent of direct-to-consumer(DTC) genetic testing [114]. The companies 23&Me and XX are perhaps the best known of these DTC options, with *Time* magazine selecting the 23&Me genetic test kit as retail invention of the year in 2008. Illumina also offers DTC genome sequencing, but this service does involve extensive involvement of a medical professional around ordering and returning results. Despite the public popularity of this kind of personalised DNA sequencing and interpretation, there are several concerns surrounding this new paradigm. The tests offered have not necessarily undergone clinical evaluation, and as they can be requested directly by the patient, the results are also interpreted by the requester without professional health-care advice. In addition, the measurable health value of the tests on offer has not been clearly established [115]. A *Lacet Oncology* editorial went to far as to call DTC tests for cancer and other conditions ”flawed and unethical” [116]. Large scale studies have begun to examine how consumers of DTC testing understand and use their genetic risk information, but the potential harms and benefits are not yet well understood.

1.9 Implementation

This thesis has largely explored exome sequencing as a proof-of-principle step in utilising the emerging genomic technology of massively parallel sequencing (MPS). However, the analysis of isolated case studies presented in this thesis has several differences to utilising this technology in a routine diagnostic service. In order to implement MPS testing in our clinical diagnostic laboratory, several issues identified in these case studies need to be considered, along with additional factors pertaining to delivering a robust clinical service which are discussed in this chapter. This is not a comprehensive summary of all the issues, requirements, and planning that is necessary, but it highlights key factors which need to be considered by the laboratory before and during the implementation process. Although general guidelines are available for clinical laboratories using MPS testing, this is a rapidly changing field and the protocols and processes constantly evolve. Therefore laboratories are required to develop their own protocols and policies around MPS testing which requires extensive consideration and planning.

1.10 Sequencing approaches

Although sequencing of the entire human genome may represent the most comprehensive test of human Mendelian disease, at the time of writing, a clinical laboratory whole genome sequencing test was not available. Sequencing of the coding regions (whole exome sequencing: WES), or disease-targeted gene panels are the main approaches used as cheaper and more practical alternatives [117]. Whole genome sequencing is beyond the scale that our clinical laboratory will likely offer in the near future and is not further discussed in depth here.

1.10.1 Disease-targeted gene panels

Disease-targeted gene panels are a logical extension of the current service of the clinical laboratory, where a physician orders specific genetic testing for a particular phenotype, and the laboratory reports a result. Disease-targeted panels contain an array of genes relevant to one or more particular clinical phenotype, which would quickly provide a result in patients that share that phenotype without the need for multiple genes being investigated in a sequential manner, and the time and money that involves [118]. These panels can be disease specific, or broader and designed to cover multiple overlapping phenotypes. A clinical application of gene panels would be those relating to specific disorders such as hereditary deafness, neonatal diabetes mellitus and Charcot-Marie-Tooth, where mutations in any number of candidate genes can be causal. Diagnosis would traditionally require sequentially screening these genes in their entirety by PCR and Sanger sequencing, a laborious and expensive process [119].

The key advantage of gene panels over exome sequencing is that targeted sequencing avoids many of the challenges described for WES, such as time, cost, reporting variants or genes of uncertain significance (VUS/GUS) or uncovering incidental findings (IF) not related to the disease phenotype being investigated [70]. Compared to exome sequencing, a limited gene set provides greater read depth in a given run, which increases analytical sensitivity and specificity, significantly increasing the confidence in variant calls. This is particularly useful in cases of mosaicism or low level heterogeneity in mitochondrial or oncology applications. One of the new challenges for laboratories developing disease-targeted gene panels is to define an appropriate list of genes that should be included for a given phenotype. This is largely due to the variability of clinical presentation and the constantly evolving understanding of which genes are disease-associated. If a laboratory chooses to offer disease-targeted gene panels, these should only include genes with sufficient evidence for a causative role in disease. In addition, any gene panel developed for a phenotype will also need to be updated with newly discovered genes with a relevance to the phenotype in question [118].

1.10.2 Exome sequencing

Although the exome makes up 1-2% of the genome, it contains >80% of the recognised pathogenic mutations. Obviously this figure represents the ascertainment bias caused by the focus on sequencing coding portions of the genome. This larger-scale sequencing moves away from the physician-ordered single gene test model and towards a hypothesis-free approach, requiring more collaboration between the laboratory and health-care provider to enable appropriate data interpretation. The current estimate for target coverage of the exome by MPS is 90-95% as certain regions are too complex for enrichment by currently available methodology. This leads to reduced analytical sensitivity (but also increased risk of incidental findings), while analytical specificity is also reduced as depth of read coverage is reduced compared to disease-targeted gene panels. One method that clinical laboratories may chose to adopt is to perform an exome sequence, but proceed only with the analysis and interpretation of know disease-associated genes, effectively creating a disease-specific analysis panel while retaining a single sample analysis work flow. Only if this does not provide a result would the remaining exome data be analysed.

Exome sequencing in the research setting has been instrumental in disease-associated gene discovery. This is now changing as more clinical labs generate exome data and are in a position to identify new causative genes, however often further functional studies are required to prove the association. The ability to perform these studies, or for an association research facility to perform them, is a new challenge that clinical laboratories are facing, as more VUS's and GUS's are identified. It may be argued that clinical laboratories are in a unique position to identify new variants that should be investigated further, particularly in cases of rare disease. Detailed clinical and phenotypic information is critical in these cases, not only for analysis and interpretation, but for longer term evaluation of variant significance. Tracking patient phenotypes and correlating them with identified genotypes can be a powerful method of evaluation that clinical laboratories could contribute to significantly. The need for periodic review and reanalysis of existing data in clinical cases is a challenge that the laboratory already faces, but the scale of this may increase with exome sequencing becoming routine practice.

1.11 Platforms

Concurrent to this PhD, our clinical laboratory was evaluating bench top NGS platforms that would be suitable to install and validate clinical testing on. At the time of starting the evaluation, two producers of sequencing platforms were considered, Illumina and IonTorrent, while multiple third parties were also considered for sample preparation, work flow automation and data analysis. While instrument specifications are easily obtained, it is difficult to maintain a current assessment of factors such as data output and run times across different platforms as each technology was being constantly improved upon and any review is rapidly outdated. The following data should, therefore, be considered as current only at the time of writing, although it will generally remain useful for comparison purposes.

1.11.1 Illumina

At the time of evaluation, llumina marketed four sequencing platforms; MiSeq, NextSeq, HiSeq and HiSeq X. These platforms differ largely in output and are designed for small and large-scale genomics. The MiSeq is the smallest platform with a maximum output of 15Gb of data and long paired-end reads of 300bp. A single run can take between 5 and 65 hours depending on the application. The NextSeq increases this maximum output to 120Gb but the paired end reads are shorter at approximately 150bp. The run time can be between 15 and 30 hours. These two instruments would be sufficient for routine diagnostic work and could handle exome sequencing or small genomes. The HiSeq was developed for large-scale genomics with a greatly increased maximum output of 1000Gb, with a paired-end read length of 150bp. Run times varied from 7 hours to 6 days for larger applications. The HiSeq X offers the greatest power for projects such as human genome sequencing. It is only sold in lots of 10, which together comprise the HiSeqX Ten platform which was the first system to achieve the $1000 human genome goal. The maximum output of this instrument was 1800Gb with 150bp paired-end reads, meaning the HiSeqX Ten is capable of producing 1.6-1.8Tb of data in a single run which takes less than three days.

1.11.2 IonTorrent

IonTorrent, was acquired by Life Technologies in 2010, which subsequently become part of ThermoFisher Scientific. Whilst still using a sequencing-by-synthesis approach similar to other systems, the IonTorrent technology measures H+ ion release during base incorporation rather than fluorescent signals. This has enabled the technology to improve rapidly as it is not relying on any optics. The fast run time and inexpensive reagents are also a positive feature of IonTorrent. The first instrument produced, the Ion Personal Genome Machine (IonPGM), was marketed as an affordable, flexible and fast sequencing system. The IonPGM can generate 200bp or 400bp reads and is compatible with three sequencing chips, the Ion314, 316 and 318 chip which greatly increases the scalability of this platform. These chips span an output range from 30Mb to 2Gb, meaning small projects can be run more economically than with other platforms in a rapid run time of 2.3 to 7.3 hours.

The IonProton was launched in sept 2012 and extended IonTorrent technology into exome and whole genome sequencing due to its larger, high density chips, allowing much larger data outputs. This instrument is compatible with the PI chip which allows an output of up to 10Gb and 200bp reads. Sequencing run times typically take between 2 and 4 hours. The Ion S5 is designed for a lower weekly throughput of gene panels, microbes, exomes and transcriptomes. It is compatible with three chips, the Ion 520, 530 and 540 chip, which all support 200bp or 400bp reads. The output of this instrument ranges from 0.6-15Gb depending on chip type and read length, though all runs can be completed in under 4 hours. The Ion S5 XL is identical in output, reads and run time to the Ion S5, but the advantage lies in the greatly reduced analysis time compared to the Ion S5.

1.11.3 Automation of template preparation

As much sample handling automation as possible is required in a diagnostic setting where large numbers of samples are processed and there is no acceptable level of mix-up. A key factor in implementing MPS involves the exploration of the various options regarding automation.

IonChef

The IonChef is an system which automates the template preparation and chip loading for the IonTorrent sequencing instruments, IonPGM and IonProton. The IonChef integrates several steps which were previously handled manually including amplification mix preparation, emulsification and amplification, enrichment, washing, sequencing prep and chip loading. The total hands on time for the IonChef is 15 minutes while the total run time will vary with chip type and read length. The IonChef is flexible in its applications so it can be used in conjunction with any IonPGM or IonProton template preparation chemistries and chips.

NeoPrep

Starting with DNA or RNA, the NeoPrep System performs all Illumina library preparation steps, including PCR amplification, quantification, and normalization. It is able to prepare 16 libraries per run, with an estimated 30 minutes hands on time. Currently the neoprep is only marketed as compatible with two library preparation methods, the TruSeq Nano DNA protocol for whole-genome sequencing applications, and the TruSeq Stranded mRNA protocol for transcriptome interrogation.

1.11.4 Third party - library prep, automation and analysis

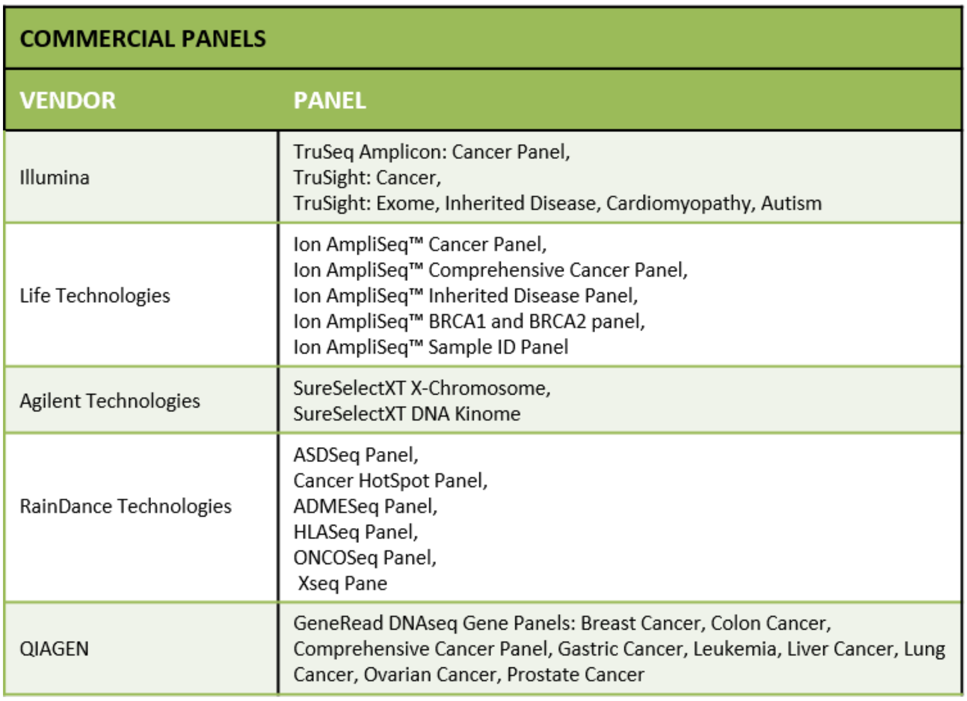
Standard enrichment kits are currently available for most key applications, including cancer, inherited disease, cardiomyopathy and autism (Table 1.1). Illumina and Life Technologies both offer products which are suitable for clinical applications, in particular the TruSight panels (Illumina) and IonAmpliSeq panels (Life Technologies). In addition to standard panels, some vendors provide custom panel design which allow a laboratory to enrich their chosen library targets.

Illumina and Life Technologies both offer custom design of their standard panels, but there are several additional custom library preparation methods provided by third parties (Table 1.2). These include Raindance (RainDance Technologies), HaloPlex and SureSelect (Agilent Techonologies), AccessArray (Fluidigm Corporation), GeneRead DNAseq (QIAGEN) and SeqCap (Roche NimbleGen) [120].

1.12 Test development

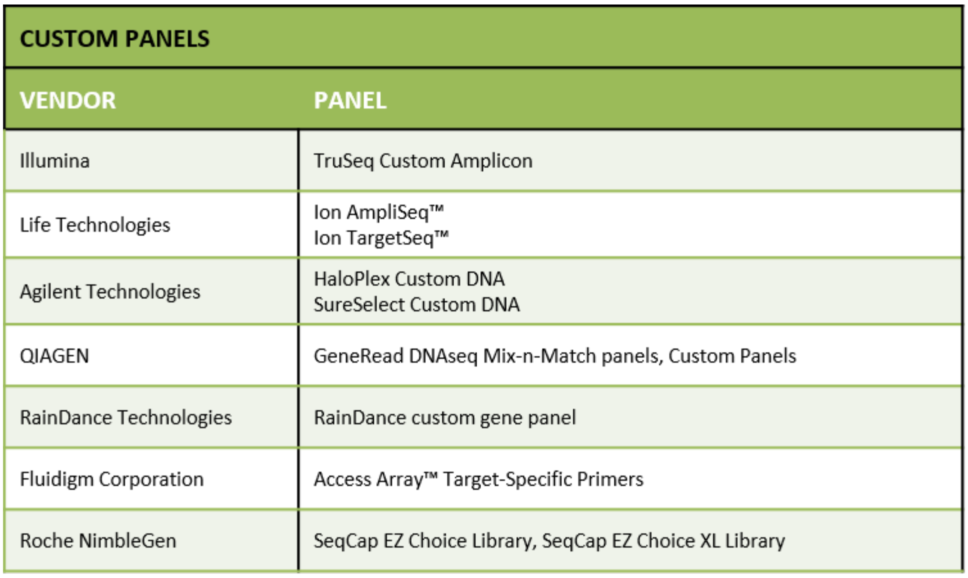
Independent of which library preparation method/target capture/NGS platform/or analytical pipeline a clinical laboratory chooses, appropriate validation of the technology must be performed during development of any clinical test. Typically each component of the sequencing pipeline will undergo cycles of rigorous performance optimisation, then the entire pipeline assessed until assay conditions and analysis settings are optimised. Part of this process is including all sample types that will be evaluated during clinical testing; whole blood, urine, saliva, FFPE. In July 2013 the American College of Medical Genetics and Genomics (ACMG) produced a document entitled *ACMG clinical laboratory standards for next-generation sequencing* [121]. This document was a guide, and it was not a requirement to follow these recommendations, although it is an excellent educational resource to guide clinical laboratories developing or validating

Table 1.1 Commercial gene panels



COMMERCIAL PANELS	
VENDOR	**PANEL**
Illumina	TruSeq Amplicon: Cancer Panel, TruSight: Cancer, TruSight: Exome, Inherited Disease, Cardiomyopathy, Autism
Life Technologies	Ion AmpliSeq™ Cancer Panel, Ion AmpliSeq™ Comprehensive Cancer Panel, Ion AmpliSeq™ Inherited Disease Panel, Ion AmpliSeq™ BRCA1 and BRCA2 panel, Ion AmpliSeq™ Sample ID Panel
Agilent Technologies	SureSelectXT X-Chromosome, SureSelectXT DNA Kinome
RainDance Technologies	ASDSeq Panel, Cancer HotSpot Panel, ADMESeq Panel, HLASeq Panel, ONCOSeq Panel, Xseq Pane
QIAGEN	GeneRead DNAseq Gene Panels: Breast Cancer, Colon Cancer, Comprehensive Cancer Panel, Gastric Cancer, Leukemia, Liver Cancer, Lung Cancer, Ovarian Cancer, Prostate Cancer

Table 1.2 Custom gene panels



CUSTOM PANELS	
VENDOR	**PANEL**
Illumina	TruSeq Custom Amplicon
Life Technologies	Ion AmpliSeq™ Ion TargetSeq™
Agilent Technologies	HaloPlex Custom DNA SureSelect Custom DNA
QIAGEN	GeneRead DNAseq Mix-n-Match panels, Custom Panels
RainDance Technologies	RainDance custom gene panel
Fluidigm Corporation	Access Array™ Target-Specific Primers
Roche NimbleGen	SeqCap EZ Choice Library, SeqCap EZ Choice XL Library

MPS testing.

Successful test development should include optimisation of factors such as the general assay conditions, read coverage, analysis settings, sample pooling and batching, costings and TAT estimates. Several of these factors impact conversely on each other (ie sample pooling and read coverage), and therefore a balance should be reached which reflects the clinical appropriateness of the testing performed. As data analysis is significantly more complex than previous single-gene sequencing assays, the performance of the sequencing pipeline should be analysed by running samples containing a range of known variants (SNVs, small indels). Recently the Genetic Testing Reference Material Coordination Program of the CDC and and the Coriell Institute for Medical Research announced the availability of transformed cell line samples for use in NGS development, validation and QC. There are DNA samples available through the Corial Cell Repositories for diseases including cystic fibrosis (CF), 5-10-methylenetetrahydrofolate reductase deficiency (MTHFR), HFE-associated hereditary hemochromatosis, Huntington diseases (HD), fragile X syndrome, alpha thalassemia and connexin 26-associated deafness. These synthetic variants may be used to optimise settings and thresholds, and compare various tools within the test pipeline. Each laboratory needs to be aware of any challenges in sequencing variants for a given disease target [122], and confirmation or gap-filling Sanger sequencing may need to be performed alongside the NGS testing in routine practise as part of the NGS testing process.

1.13 Validation

Following development and optimisation of pipeline conditions, validation of the entire test should be performed in an end-to-end manner. Assay performance should be assessed for analytical sensitivity (false negative rate), specificity (false positive rate), repeatability and reproducibility. Generally the first test developed by the laboratory will undergo very extensive validation, while subsequent tests that modify the pipeline slightly may only require validation of these changes. Although each test will require their own particular validation focus, in general, every test needs to pass minimum metrics that define a high-quality run. These metrics will need to be defined by the laboratory and should cover such factors as: minimum read coverage, average coverage across genome, percentage of target area covered to list a few. In addition, the accuracy of sequencing across challenging regions such as homopolymer stretches, or regions of high homology should be assessed. For disease-targeted panels, validation should include assessment of all common mutations that should be detected in clinical samples. CDC-Corial Cell Repository samples, previously-tested patient samples, or other synthetic controls which are well-characterised reference samples may be used to extend the spectrum of variants available to validate. Exome sequencing validation relies more on the metrics of a high quality run, in addition to analysis of samples that have been previously analysed using an alternative platform. Validation samples that are being run through the entire pipeline should be blinded to ensure quality of the validation process.

1.14 Quality Control and Proficiency Testing

Each laboratory should develop quality control measures which can be applied to every MPS run. Typically these measures act to identify sample-prep failures or sequencing run failures so the next processing step is not initiated prior to samples passing these checks. In addition, sample identity must be accurately tracked throughout the testing process which is particularly critical when barcoded samples may be pooled. Proficiency testing should be carried out periodically, and the laboratory should investigate suitable schemes to enrol in based on what testing is being offered, as well as internationally accredited auditing processes, although generally there is no external quality assessment or reference method for MPS.

1.15 Technical considerations

Many file types exist for exporting raw variants and annotation data though the more commonly used are bam files for alignments and fastq files for sequence reads. A standard file format that has emerged is the variant call format (vcf). The Centres for Disease Control and Prevention is currently leading an effort to develop a consensus gVCF file format. This format is designed to store sequencing information for both variant and non-variant positions which is required for human clinical applications and involves a set of conventions applied to the standard variant call format as documented by the 1000 genomes project. Many of the issues with file type compatibility and conversion were present in the early days of MPS, and most platforms and software now handle the standard file types with ease.

The biggest challenge that MPS presents to the clinical laboratory infrastructure is that of data storage. In addition to the huge amount of data being produced on instrument, there are also large databases to be accessed and updated. The two issues the laboratory will face are the immediate storage of data and transmission of data to and from databases or other users. An on-site server allows for the storage of much greater data volumes, and is a near essential component for a diagnostic lab considering MPS technology on a larger-scale. There appears to be a movement within the bioinformatic community towards cloud computing. Described in 2010 by Stein *et al* [123] as a general term for computation-as-a-service; cloud computing allows users to pay for the usage of hardware and storage for as long as they require for their purposes. Each laboratory will need to consider their own data storage needs, and data security requirements to decide on an appropriate solution to this emerging issue.

1.16 Data analysis

One of the biggest challenges in MPS is the hugely increased amount of data produced per sample which needs to be accurately analysed. The ACMG has published a specific set of guidelines entitled *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology* [124], which provide guidance for the clinical laboratory on the various approaches to variant classification.

1.16.1 Incidental Findings

As with all genetic testing, appropriate informed consent is important to ensure the patient is aware of the potential harms that may occur as a consequence of the testing process. Any comprehensive test, including arrayCGH and microarray, may result in findings unrelated to the disease status of the patient, but which may have other health implications now or in the future. These findings, commonly referred to as incidental findings (IF), are significantly increased with MPS methods and have been a source of concern since the early implementation of MPS. No specific guidelines exist for appropriate consenting, and each institution offering MPS testing are currently devising their own policies, particularly concerning the level of detail provided in reports. The ACMG does provide recommendations for reporting incidental findings. They state that a laboratory should provide the following information about incidental findings: (i) Whether the laboratory will systematically search for and report on the ACMG recommended variants (detailed in the *ACMG Recommendations for reporting of Incidental Findings in Clinical Exome and Genome Sequencing* [125]; (ii) whether IFs are routinely confirmed to ensure analytical accuracy, or whether confirmation is recommended through additional follow-up testing; (iii) a clear definition of the criteria used to decide which types of IF will be reported and how these can be requested or declined. A previous version of the ACMG guidelines recommended that there be an opt-out process for patients who did not wish to receive IF results, however this has been amended to a decision which is made in the consenting process, so the ordering provider and patient are clear on what data will be disclosed prior to the testing being performed.

1.16.2 Finding of uncertain significance

Similar challenges apply to identified variants which may be related to the disease phenotype in a patient, but for which little or no classification is available. This is not a situation unique to genetics, as all medical tests can highlight abnormalities that were incidental to the the main line of enquiry. In terms of MPS these are termed variants of uncertain significance (VUS), and often occur in genes of uncertain significance (GUS) which have never been associated with any patient phenotype previously, or have only been associated with a phenotype not observed in the patient in question. Although VUS are more common in exome or genome sequencing, some laboratories have also acknowledged the issue in larger gene panels which cover multiple phenotypes. A response to this is to allow physicians the choice of requesting that only the relevant subset of genes will be analysed. VUS's that are discovered as IF's in genes unrelated to the diagnostic evaluation should not be returned because they have no clinical implications and are more likely to cause confusion than harm. VUS may be considered as candidates for disease causality, but must be reported as of uncertain significance until additional evidence is obtained that support an association to disease. Multiple cases with deleterious variants in the same gene would provide evidence for association and allow the variant to be re-classified according to the specific recommendations of the ACMG guidelines (Pathogenic/Likely pathogenic/Benign/Likely Benign). This raises interesting questions about the responsibility of the laboratory to review historical VUS cases in light of new evidence for disease association, this is particularly pertinent for exome and genome sequencing. Should the laboratory undergo periodic review of VUS in exome and genome cases where a cause of disease was not found, or should a clinical report include advice to the requesting professional that they request re-analysis of the data at a future date? Additionally, is there any responsibility on the laboratory to contribute to research aimed at providing evidence on potential pathogenicity of VUS, either by aligning with a research laboratory or curating an internal database of phenotype/genotype data? As shown in the case presented in Chapter 5, advances in knowledge on VUS's can be a long and slow process. It is recommended that the laboratory carefully develop a policy and process for returning information and ensuring that it conforms to accepted medical and ethical obligations.

1.16.3 Computational predictive programs

Computational predictive programs (*in silico* analysis) can be used to aid the interpretation of sequence variants. There are many tools available, but generally the algorithms used determine the effect of the variants at the nucleotide and amino acid level, and the impacts that then may have on protein function. An important factor to consider when using computational predictive programs is that these tools are between 65 and 80% accurate [126] even when used to predict clinically proven disease-causing variants. Generally these tools have a low specificity, resulting in over-classifying missense changes as deleterious [127]. Due to this inaccuracy, it is not recommended that these predictions be used as the single source of evidence when classifying a VUS.

1.17 Reporting

There are several guidelines on the essential elements of a clinical laboratory report [110, 121], which all emphasise the importance of succinct, informative reports which are nevertheless easy to understand. Reports should include structured results, an interpretation of the results, methodology, references and disclaimers.

Results

The results section of a clinical report should clearly state the variants identified that pass all validation and QC metrics, as well as the ACMG classification that these fall into. In the instance of larger gene panels, or exome and genome sequencing, this list may be extensive and may be more clearly presented in table form. Nomenclature of variants should be in accordance with the standards of the Human Genome Variation Society (HGVS) [128]. Laboratories should report the version of nomenclature used in their test methods. Sequence references should be provided to ensure unambiguous identification of the variant. Genome sequence ”g.”, coding sequence ”c.” and protein sequence ”p.” should be used to assist in functional interpretations. For mitochondrial variants ”m.” should be used. If historical nomenclature for a variant differs from the current nomenclature, an additional notation should draw attention to this. The reference sequence used for each test should be derived from either the National Centre for Biotechnology Information (NCBI) RefSeq database (http://www.ncbi/nlm.nigh.gov/RefSeq) or the of the Locus Reference Genomic database (http://www.lrg-sequence.org). If using the NCBI reference, the version number must be recorded.

Interpretation

The interpretation of the result provides evidence supporting the variant classification that has been made, and identifying the impact of these variants on the patients phenotype. The interpretation should also include recommendations for any supplementary testing which could further support the variant classification.

Methodology

This section should detail the test method; Disease-specific panel/Exome for example, and what types of variants are expected to be detected. It is important for laboratories to be aware of refractory regions in any commercial enrichment protocol used in exomes/gene panels (genomes don't need enrichment), and report the method used and the capture or enrichment efficiency expected based on their validation studies.

1.18 Thesis direction

DNA sequencing is a core technology that over the last 40 years has contributed greatly to biology as a whole, and in particular the characterization of the genome, although historically it did not drive diagnosis so much as have a confirmatory role. The recent advent of massively parallel sequencing (MPS) technologies has revolutionised the role of DNA sequencing in both biological research and human diagnostic genetics settings. Indeed, the new technology has enabled sequencing techniques to be used in a much more diverse range of disciplines, including for previously unconsidered applications [129].

This thesis considers the developing state of genomic technologies and explores how these can be applied within the sphere of clinical diagnostic molecular genetics. In particular it focuses on the role of exome sequencing in rare disease, as this approach encompasses many of the processes and issues discussed in this introductory chapter and is an approach currently being successfully applied by many clinical laboratories to provide a genetic diagnosis in complex cases.

The number of rare diseases described to date lies between 5000 and 8000 [5]. Although the definition of 'rare' varies between countries, the EU definition of a rare disease affects less than 5 individuals in 10,000 [130], collectively affecting an estimated 29 million people [131]. In the US, rare disease affects up to 25% of paediatric inpatient admissions [132], though unfortunately, only 46% of patients referred to a medical geneticist will receive a genetic diagnosis [133]. A genetic diagnosis can affect both patient treatment and management, as well as further or future reproductive choices where prenatal, or pre-implantation (PGD) screening may be possible. Thus a diagnosis can benefit the affected individual and the wider family even where there is no known treatment.

This body of work will focus on investigations in three families, where one or more individuals have a clearly inherited disease or syndrome, but traditional diagnostic techniques have failed to identify a genetic cause. The following chapters will test whether we can implement WES in a clinical setting for complex cases, and highlight difficulties associated with this technique, with the ultimate aim being the introduction of MPS into the NZ diagnostic setting. Given the requirement for strict auditing of analysis and reporting, and the drive for maximal automation of diagnostic processes, this thesis focusses on commercial solutions for data handling. Supplier analysis solutions have been preferred over third party commercial software to ensure that validation of a pipeline or process would not need to be repeated if a third party supplier stop supporting a software. Additionally, given that the work encompassed in this thesis was performed in a clinical diagnostic setting, this impacted on the selection on case individuals and the resources to follow up on functional validation of candidate variants. The specific aims of this thesis are outlined in Table 1.3.

Table 1.3

	Thesis Aims and Objectives
1	Explore the benefits and challenges of NGS technology in complex diagnostic cases.
2	Analyse exome sequencing data in multiple case studies with the aim of identifying candidate gene variants.
3	Provide recommendations on implementing NGS technology into the diagnostic laboratory.

Chapter 2

Materials and Methods

2.1 Suppliers and nomenclature

2.1.1 Equipment, reagents and consumables

The source details of each item is given in full in the first instance in the chapter, after which the item is referred to by name only.

2.1.2 Nomenclature

The nomenclature for DNA and protein mutations used in this thesis aligns to the recommendations of the Human Genome Variation Society (www.hgvs.org/mutnomen/) unless otherwise stated.

2.2 Clinical Samples

Patient samples were obtained from referring physicians who obtained the requisite consent for diagnostic testing. Several sample types were collected, including peripheral blood collected by venipuncture into 10ml EDTA tubes, formalin-fixed parafen-embedded (FFPE), post-mortem tissue and blood spot cards.

2.3 General Methods

2.3.1 DNA extraction

Genomic DNA extraction (Manual method)

Genomic DNA (gDNA) was extracted from peripheral blood collected into 5ml EDTA vacutainer tubes using an adapted guanidine thiocyanate/isopropanol method. EDTA blood was mixed 1:4 to lysis buffer (0.32M sucrose, 10mM HEPES (pH7.5), 5mM MgCl2, 1% (v/v) Triton X-100, 0.1% (w/v) sodium azide) and mixed with gentle inversion to lyse red blood cells. Leukocytes were then pelleted by centrifugation at 3000 x g for 10 min and the supernatant decanted and discarded. The nuclear pellet was resuspended in 5mL of denatuation solution (4M guanidine thiocyanate, 25mM sodium acetate, 0.84% (v/v) beta-mercaptoethanol) by vigorous vortexing and the tube placed on a roller and mixed overnight. An equal volume of isopropanol was added and the tube mixed by gentle partial inversion to precipitate the gDNA. The filamentous DNA precipitate was collected by centrifugation, spooled out and washed twice in 70% (v/v) ethanol before being pelleted by a pulse centrifugation and the excess ethanol carefully removed. The precipitated DNA was resuspended in 0.2-0.5mL of 0.1 x Tris EDTA buffer (1mM Tris-HCL (pH8.0), 0.1mM EDTA) at 4°C for ∼24 hours before being placed on a rotating wheel at room temperature for a further 48 hours to ensure complete resuspension.

Genomic DNA extraction (Spin column method)

Genomic DNA (gDNA) was extracted from peripheral blood collected into 5ml EDTA vacutainer tubes using the Machery-Nagel NucleoSpin Blood-L (Machery-Nagel, Duren, Germany) method. 2mL of whole blood was added to 150µL proteinase K and 2ml of Buffer BQ1 and vortexed thoroughly.

Blood was incubated at 56°C for 30 min (vortexing every 10 min) before being cooled to room temperature. 2mL of 100% (v/v) ethanol was added and mixed immediately by inversion. The lysate was pipetted into a Nucleospin filter tube (3mL at a time) and centrifuged for 3 min at 5000 rpm. The flowthrough was discarded and 2mL of Buffer BQ2 added to the tube.

This was centrifuged for 2 mins at 5000 rpm and a further 2mL of Buffer BQ2 added. Tubes were centrifuged for 10 min at 5000 rpm and the column placed into a fresh collection tbe. A volume of 200µL of preheated (70°C) buffer BE was added directly to the silica membrane and incubated at room temperature for 2 min. A final centrifugation at 5000 rpm colleted the eluted DNA into the sample collection tube.

DNA extraction from non-blood samples

Genomic DNA (gDNA) from bucal cells, fresh tissue, FFPE tissue and cultured cells was extracted as above using the Machery-Nagel method with the variations detailed in the kit instructions for non-blood samples.

DNA extraction from Guthrie card

Genomic DNA (gDNA) was extracted from Guthrie card blood spots using an in-house extraction method. For each patient, a 3mm punch from a dry Guthrie spot was placed into a 600µL 96-well plate. 60µL of 50mM NaOH was added to the well containing the punch and the plate incubated at 100°C for 15 mins. Following incubation, the plate was removed from heat and 15µL of 1M HEPES (pH7.5) was added to the sample. The plate was centrifuged at 14000rpm for 30s to pellet the blood spot and particulates. The resulting supernatant is the DNA stock solution used directly in subsequent PCR reactions. No quantification is performed on this DNA as it is a crude extraction and concentration is typically $<5\text{ng}.\mu\text{L}^{-1}$. This was sufficient for standard PCR reactions.

2.3.2 DNA quantification

Aliquots of 2µL gDNA were analysed by the ND-1000 UV-visible spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). DNA concentration was measured by absorbance at 260nm using the conversion coefficient; $1\text{AU}_{260}=50\text{ng}.\mu\text{L}^{-1}$

2.3.3 Standard Polymerase Chain Reaction

Standard polymerase chain reaction (PCR) amplification was performed with FastStart Taq DNA polymerase (Roche) using the buffer and reagents provided. Oligonucleotide primers were designed manually and ordered from IDT (Intergrated DNA Technologies Inc., Coralville, IA, USA) Standard reactions contained 200µM of each deoxribonucleotide triphosphate (dNTP), 1x reaction buffer, 0.5U of FastStart Taq DNA polymerase, 0.25µM of each oligonucleotide primer (forward and reverse), 20ng gDNA and sterile water to a final volume of 10µL. Reactions were prepared in 0.2mL thin wall polypropylene tubes (Axygen Scientific Inc., Union City, CA, USA) and thermal cycling was performed on an iCycler (Bio-Rad Laboratories Inc., Hercules, CA, USA) or a SimpliAmp (Applied Biosystems Corp., Foster city, CA, USA) instrument. Thermocycling parameters were as folows: Annealing temperatures (T_A) were determined empirically for each primer pair.

96°C	4 mins	
96°C	30s	
T_A	30s	
72°C	1 min	29x
4°C	Hold	

Oligonucleotide primers were designed to be approximately twently nucleotides in length, with approximately 50% guanine/cytosine composition (G/C%). Negative (no DNA) amplifications were run in parallel, and where possible, positive controls were included.

2.3.4 Agarose gel electrophoresis

Small (6.0 x 8.5cm) gels were cast from low electroendosmosis (LE) agarose (Roche Diagnostics GmBH, Mannheim, Germany) prepared at 0.7% or 2% (w/v) in Tris-borate EDTA buffer (TBE; 90mM Tris-borate, 2mM EDTA) with CyberSafe interchelating dye. Combs were inserted at the anodal end and the gel set by cooling to RT. Once set, gels were immersed in 0.5x TBE and the comb removed. For assesing DNA quality, 2-5µg of whole gDNA was digested, then mixed with 5µL of a 5 x loading buffer (40% (w/v) sucrose, 0.25% (w/v) bromophenol blue (BPB), 0.25% (w/v) xylene cyanol (XC)) and loaded onto a 0.7% gel with a 12Kb DNA ladder.

When assessing PCR reactions, 2µL of PCR product was mixed with 2µL of 5x loading buffer buffer (40% (w/v) sucrose, 0.25% (w/v) bromophenol blue (BPB), 0.25% (w/v) xylene cyanol (XC)) and loaded onto a 2% gel. 2.5µL of either *Hind*III digested PhiX DNA [10ng.µL^{-1}] or 100bp DNA ladder [50ng.µL^{-1}] (Both from Roche Applied Sciences) were loaded to provide size standards. Samples were electrophoresed at 1-2 V.(cm^2)$^{-1}$ until the BPB dye front had sufficiently migrated toward the cathodal end of the gel. Visualisation and capture of digital images were performed with a Gel Doc 2000 molecular imager and Quantity One v.4.6.5 software (Bio-Rad Laboratories Inc.).

2.3.5 CGH Array

Array CGH to determine copy number variants was performed using Nimblegen CGX12 arrays Roche NimbleGen, Mdison, WI, USA), containing 135,000 oligos across the genome. Analysis of array data was performed using Genoglyphix v3.0-0 software (PerkinElmer, Waltham, MA, USA).

2.3.6 Automated DNA sequencing methods (Sanger)

Sequencing reactions

PCR products were initially diluted 1:4 in sterile water prior to inclusion in a sequencing reaction. Sequencing reactions were prepared with 500pmol of oligonucleotide primer, 20ng of purified PCR product, 0.5µL BigDye Terminator v.3.1 enzyme (Applied Biosystems) and and 2µL of 5x sequencing buffer (Applied Biosystems). Reactions were prepared in 0.2mL thin wall polypropylene tubes (Axygen Scientific Inc., Union City, CA, USA) and thermal cycling was performed on an iCycler (Bio-Rad Laboratories Inc., Hercules, CA, USA) or a SimpliAmp (details needed) instrument. Thermocycling parameters were as folows:

2.3.7 Sequencing reaction purification

Sequencing reactions were firstly diluted 1:4 in sterile water and then purified by resin filtration. 800µL of a Sephadex G-50 fine DNA grade resin (GE Healthcare, Chalfont St. Giles, UK)

96°C	10 mins	
96°C	10s	
50°C	10s	
60°C	2 min	25x
10°C	Hold	

(hydrated in sterile water) was pipetted into the wells of a Unifilter 800 96-well microtitre plate. Excess sterile water was removed by centrifugation at 750 x g for 5 min and the Unifilter plate placed on a semi-skirted 96-well PCR plate (Scientific Specialties Inc., Lodi, CA, USA) containing 20µL of sterile water in appropriate wells. 10µL of sequencing reactions were transferred onto the sephadex column and centrifuged at 750 x g for 5 min.

2.3.8 Sequencing analysis

Sequencing was perfomed on a 3130XL Genetic Analyser (Applied Biosystems) using POP-7 polymer (Applied Biosystems) using a 50cm, 16-capillary array. The data was collected using Data Collection software v.3.0 as per the manufacturers instructions (Applied Biosystems). Sequence data was analysed using Sequencing Analsyis software v.5.2 and Geneious Pro software v.5.6.4 (Biomatters Limited).

2.4 Methods used in Chapter 3 Analysis of an intellectual disability pedigree

2.4.1 Ion Proton sequencing methods

Library preparation

Exome library preparation was performed using the Ion Ampliseq Exome RDY Library Preparation kit (Catalog # 4489840). For each sample a master mix was prepared containing 14µL 5 x Ion AmpliseqTM HiFi Mix, 50-100ng gDNA, and sterile water to make up to 70µL total volume. A volume of 5µL was dispensed into 12 wells of an Ion AmpliSeq Exome RDY plate which contains the dry primers, before the plate was sealed and run on the following thermocycler program. The 12 PCR reactions for each sample were combined into a central well in the plate. 6µL of FuPa Reagent was added to this well to partially digest primer sequences. The plate was sealed, vortexed, and spun down to remove droplets before being run on the following thermocycler program.

99°C	2 mins	
99°C	15s	
60°C	16 mins	10x
72°C	30s	

50°C	10 mins
55°C	10 mins
60°C	20 mins
10°C	Hold

Ligation of barcode adapters

In preparation for running pooled exome libraries on one P1 chip, Ion P1 Adapters and Ion XpressTM barcodes were added. For each sample a mix was prepared with 1.5µL Ion P1 Adapter, 1.5µL Ion XpressTM Barcode X (unique for each mix) and 3µL nuclease-free water. 6µL of this mix was pipetted into the plate well containing digested amplicons. 12µL of Switch Solution and 6µL of DNA Ligase is added to bring the total reaction volume to 80µL. The plate was sealed, vortexed, and spun down to remove droplets before being run on the following thermorcycler program.

22°C	30 mins
72°C	10 mins
10°C	Hold

Purification of library

80µL (1 x sample volume) of Agencourt AMPure XP Reagent (thawed slowly to room temperature and vortexed before use) was added to each library and the bead suspension mixed thoroughly by pipetting up and down 5 times. The bead/DNA mix was incubated for 5 min at room temperature and then for 2 min in a magnetic rack until the solution cleared. The supernatant was carefully removed and discarded. 150µL of freshly prepared 70% ethanol was added to wash the beads, then removed and discarded. This wash was repeated a second time. The beads were air dried in the magnet at room temperature for 5 min.

Quantification of library by qPCR

The plate was removed from the magnet and 50µL of Low TE was added to the pellet to disperse the beads. The plate was sealed, vortexed, and spun down to remove droplets before being placed back in the magnet for >2 min. A 100-fold dilution of the supernatant was prepared for quantitation with the Ion Library Quantitation kit. Three 10-fold serial dillutions of the E. coli DH10B Ion Control Library (~68pM) were prepared at 6.8pM, 0.68pM and 0.068pM. For each sample, standard and negative control the reaction mixture prepared contained 20µL of 2 x TaqMan MasterMix and 2µL of 20 x Ion TaqMan Assay. 11µL of this mix was aliquoted into a 96 well PCR plate. 9µL of the diluted (1:100) library or control dilution was added to appropriate wells (in duplicate reactions). The final reaction volume for each duplicate was 20µL. The qPCR instrument was programmed with the concentration of the control library standards. ROXTM Reference Dye was the passive reference dye used, and FAMTM dye/MGB was the TaqMan probe reporter/quencher option selected. The qPCR parameters are as follows. Based on the the calculated library concentrations, libraries were combined (two per pool) so that the resulting sample contained equal pM of library. The final concentration aimed for was

50°C	2 mins	
95°C	20s	
95°C	1s	
65°C	20s	40x

100pM, although this was lowered to 50pM for some samples.

Template preperation Ion ChefTM system

Pooled libraries were loaded onto the Ion P1 chip using the Ion ChefTM system with the Ion P1TM 200 Kit (Catalog number 4488377). The protocol for this stage is detailed in the Life Technologies user guide (publication number MAN0010078).

2.4.2 Ion ProtonTM Sequencing

Sequencing of exome libraries was performed on the Ion ProtonTM instrument as per the protocol detailed in the Life Technologies user guide (publication number MAN0010078).

2.4.3 Variant Analysis

Variant annotation and filtering was performed in IonReporterTM v.4.2 (Thermo Scientific).

Variant Analysis - Proband only

Analysis was initially performed on exome data from the proband only. The filtering parameters employed in IonTorrent are described in Table 2.1

Variant Analysis - Multiple siblings

Analysis was subsequently extended to three siblings (proband included) and widened to include autosomes. The filtering parameters employed in IonTorrent are described in Table 2.2

2.4.4 Resequencing

DACH2

Resequencing to confirm the DACH2 mutation detected in exome data was performed under standard conditions (see 2.3.3). Sequences of oligonuclotide primers used are detailed in Appendix A.2.

Table 2.1 Proband-only X chromosome variant filtering

Total Variants:759		**Total Genes:10263**
10<=Filtered coverage<=100000		
	Variants:648	
Zygosity in Homozygous		
	Variants:530	
Location in Exonic		
	Variants:239	
0.0<=SIFT<=0.2 OR 0.8<=PolyPhen<=1.0 OR Grantham <=215.0		
	Variants:64	
Variant Effect in: Stoploss, nonsense, missense, frameshift insertion, nonframeshift insertion, framesift deletion, nonframeshift block substitution, nonframeshift deletion, frameshift block substitution		
	Variants:64	
0.0<=Minor Allele Frequency<=0.02		
	Variants:3	**Genes:3**

Table 2.2 Sibling variant filtering

Total Variants:86914		**Total Genes:24848**
Variant Effect in: Nonsense, missense,		
0.0<=SIFT<=0.2 AND 0.8<=PolyPhen<=1.0 0.0<=Minor Allele Frequency<=0.01 20<=Filtered coverage<=100000		
Common to all three siblings		
	Variants:66	**Genes:58**

KMT2C

Resequencing of the KMT2C mutations detected in exome data was performed under standard conditions (see 2.3.3). Sequences of oligonuclotide primers used are detailed in Appendix A.2. Due to suspected non-specific amplification of KMT2C, PrimerBlast tool was used to identify a BAGE gene family member on chromosome 21 which was an unintended target of the orginal primer pairs. Redesign of new allele specific primers was performed using the Web-based Allele Specific Primer (WASP) tool described by Wangkumhang P *et al* [134]. Sequences of oligonuclotide primers used are detailed in Appendix A.3.

2.5 Methods used in Chapter 4 Trio analysis - a sporadic case of premature aging

IonTorrent sequencing method previously described in 2.4.1.

2.5.1 Variant Analysis

Variant annotation and filtering was performed in IonReporterTM v.4.2 (Thermo Scientific).

Recessive inheritance

An initial filtering approach was used to identify compounding or homozygous mutations that would suggest a recessive disease inheritance mode. The filtering parameters employed in IonTorrent are described in Table 2.3

De novo inheritance

An second approach aimed to identify new mutations present only in the proband that would suggest a *de novo* inheritance mode. The filtering parameters employed in IonTorrent are described in Table 2.4

Table 2.3 Recessive analysis filtering pipeline

Total Variants:72545		**Total Genes:16205**
0.0<Minor Allele Frequency<0.01		
	Variants:11358	
Variant Effect in Stoploss, nonsense, missense, frameshift insertion, nonframeshift insertion, framesift deletion, nonframeshift block substitution, nonframeshift deletion frameshiftblock substitution		
	Variants:2557	**Genes:1171**
Genetic Category Type in In transphase compound heterozygote		
	Variants:154	**Genes:118**
0.0<SIFT<0.2 AND 0.8<PolyPhen<1.0		
	Variants:33	**Genes:33**
UCSC Common SNPs OUT		
	Variants:1	**Gene:1**

2.5.2 Resequencing

BCOR

Resequencing to confirm the BCOR mutation detected in exome data was performed under standard conditions (see 2.3.3). A nested allele-specific PCR approach was employed to investigate parental mosaicism, using allele specifc primers in a second round of standard amplification. Sequences of oligonuclotide primers used are detailed in Appendix A.3.

2.5.3 Allele specific amplification of BCOR

In order to try and detect parental mosaicism of the BCOR mutation identified in the affected child, an allele-specific PCR assay was designed with primers targeted specifically to the wild-type (wt) or mutant allele. Design of the specific oligonuclotide primers was performed using

Table 2.4 De novo filtering pipeline

Total Variants:72545		**Total Genes:16205**
0.0<Minor Allele Frequency<0.01		
	Variants:11358	
Variant Effect in Stoploss, nonsense, missense, frameshift insertion, nonframeshift insertion, framesift deletion, nonframeshift block substitution, nonframeshift deletion frameshiftblock substitution		
	Variants:1613	**Genes:1336**
Genetic Category Type in HasDeNovoNonRefAllele		
	Variants:3	**Genes:3**

the Web-based Allele Specific Primer (WASP) tool [134]. Sequences of oligonuclotide primers used are detailed in Appendix A.3.

2.5.4 X-inactivation analysis

Digestion and PCR

As part of investigations into X-linked gene mutations, an X-inactivation assay was adapted to determine if non-random X-chromosome inactivation could be observed in extracted DNA from various tissues. The final protocol was a modification of the protocol described in Current Protocols in Human Genetics (Nonrandom X Chromosome Inactivation Detection : Supplement 80). The assay is based on a methylation-sensitive restriction digestion with *Hpa*II which will only cleave unmethylated DNA, which in this situation equates to the active X chromosome. The inactive X chromosome will be methylated and therefore not a target of *HpaII.* A PCR of the highly polymorphic CAG repeat in the coding region of the human androgen receptor gene(*AR*) is performed on the predigested samples and the methylation status of the X chromosomes can be determined by comparison of the digested/undigested sample on fragment analysis.

Two seperate restriction predigestion reactions were prepared for each patient DNA sample, one with *Hpa*II and one without. In addition to the patient samples, a previously determined highly skewed sample, a healthy female sample, and a no-template sample were included as controls. The two reaction mixtures were prepared with 1µL 10x Cutsmart buffer, 2µL100ng/µL DNA. The enzyme-included digestion reaction included 1µL10 U/µL of *Hpa*II. Both reaction mixtures were made up to 10U/µL with sterile water. The no-template sample was prepared with 2U/µL sterile water in place of DNA. The reactions were mixed gently and incubated at 37°C for >60 mins - not exceeding 16 hours. After incubation the samples were incubated at 80°C for 20 mins to deactivate the *Hpa*II.

One PCR mixture was prepared and amplification performed on the both the digested and undigested samples. The PCR mixture contained 2.5µL 10x AmpliTaq Gold PCR buffer (without magnesium), 1µL25 mM $MgCl_2$, 5µL dNTP mix (0.25mM each), 1.5µL DMSO, 0.5µLAmpliTaq Gold Polymerase (5U/µL), 10 µM primer mix (primers detailed in Appendix A.3) and 10µL sterile water to a total volume of 22.5µL/reaction. 2.5µL of pre-digested/undigested template was added to the PCR mix and run on the following thermocycler program.

94°C	10 mins	
94°C	30s	
62°C	30s	
72°C	30s	30x
10°C	Hold	

Fragment analysis

The products of this reaction were diluted 1:5 in sterile water and 1µL of diluted template was combined with 0.2µL LIZ 500 size standard (Thermo Fisher Scientific, Waltham, MA, USA) and 12µL of Hi-Di formamide (Thermo Fisher Scientific, Waltham, MA, USA) per reaction in preparation for fragment analysis. Each sample was prepared in duplicate and denatured for 2 mins at 95°C. Capillary electrophoreses was performed on the ABI 31300XL per manufacturer instructions. Raw data was processed with the GeneMapper software to determine the peak area for each amplified allele. The calculation of the relative X-inactivation ratio for each sample was performed as per the original protocol.

2.6 Methods used in Chapter 5 A case of sudden cardiac death

2.6.1 MiSeq sequencing methods

Library preparation

Library preparation was performed with the Illumina TruSeq DNA LT sample Prep Kit, using the protocol detailed in the TruSeq DNA Sample Preparation Guide (Part # 15026486 Rev.C).

Exome enrichment

Exome capture (enrichment) was performed with the Illumina TruSeq Exome Enrichment Kit, using the protocol detailed in the TruSeq Enrichment Guide (Part # 15013230 Rev. J).

2.6.2 Sequencing protocol

Exome sequencing was performed on the Illumina HiSeq 2000 (paired-end 100 base pair reads) using the protocol detailed in the HiSeq 2000 System User Guide (Part # 15011190 Rev. M).

2.6.3 Variant annotation

Initial data analysis was performed with the Illumina HiSeq analysis software enrichment pipeline v.0.9 using default parameters. The pipeline includes Burrows-Wheeler Aligner v.0.6.1-r104-tpx, and Genome Analysis Toolkit (GATK) v1.6-22-g3ec78bd.

2.6.4 Variant filtering

Variant calls were further annotated and filtered with ANNOVAR and ENCODE Gencode v.12 (GRCh37). Variants were sequentially filtered against the 1000 Genomes (Apr 2012 release) and dbSNP137 databases, whereby variants with a minor allele frequency >0.01 were removed. Synonymous variants were excluded from analysis. Annotated variants were initially filtered on segregation with the phenotype. Only variants shared by the two affected individuals were retained. Two independant lines of further filtering were then performed on this subset:

Candidate gene analysis

A list of known cardiomyopathy genes was then interrogated in this subset, and any genes represented in the affected individuals (as containing shared variants) were marked for further investigation.

Exome wide analysis

A filtering system based on a recessive inheritance pattern pulled out any genes in which there were two or more shared variants in the two affected individuals. These genes were then systematically reviewed for variant phasing in the unaffected parents, and potential association with phenotype. Variants in genes of interest that passed this filter were marked for confirmation and further analysis by Sanger sequencing in additional family members.

2.6.5 Resequencing

Resequencing to confirm PPA2 mutations detected in exome data was performed under standard conditions (see 2.3.3). Sequences of oligonuclotide primers used are detailed in Appendix A.4.

2.6.6 Structural modelling

Structural modelling of human PPA2 in order to determine positioning of mutant residues was based on the high homology between human and yeast ppa2. Basic space fill models of ppa2 were visualised in Cn3D.

Modelling of PPA2 and MT- COX IIII was performed using Swiss-Pdb viewer (Spdbv)/ DeepView for OSX v.4.1.0 (www.expase.org/spdbv/). Reference pdb files were sourced from the RCSB Protein Data Base (www.rcsb.org/pdb/home/home.do). PPA2 base model ID: PDB=1M38_B/

MMDB=21720 (Saccharomyces cerevisiae), MT-COXIII base model ID: PDB=2OCC_C/MMDB=53955 (Bovine).

2.7 Methods used in Chapter 6 Wider investigations of PPA2

2.7.1 Structural modelling

Software used to produce structural models is previously described in 2.6.6.

2.7.2 Oxidative stress sensitivity of PPA2 knockout yeast

All yeast methods were performed in conjunction with Thomas Hoefken (Brunel University, Uxbridge, London UK), and relied on his expertise and laboratory setup. Oxidative stress sensitivity of *ppa2* knockout yeast was assessed by comparative growth to WT yeast on media containing diamide, which specifically oxidises intracellular thiols. Plates were prepared containing diamide at a concentration of 0mM, 1mM, 1,25mM and 1.5mM. A volume of 4µL of *ppa2* knockout and WT yeast cell dilutions (at an OD_{600} of 0.1, serially diluted to 1:10, 1:100 and 1:1000) were spotted onto plates at each diamide concentration and incubated at 37°C for 2-3 days to allow sufficient growth for comparison of WT and *ppa2* knockout strains in each condition. Plates were photographed using a Geldoc system under white light.

2.8 Methods used in Chapter 7 Zebrafish modellling of PPA2 dysfunction and mitochondrial disease

2.8.1 Zebrafish husbandry

Zebrafish used in this reseach were contained in the the Otago Zebrafish Facility (Department of Pathology, Dunedin School of Medicine, Dunedin, New Zealand). Fish procedures and husbandry protocols are detailed in the Otago Zebrafish Facility Standard Operating Procedures. Zebrafish tanks were connected in a self-siphoning ZebTec system (Techniplast, Italy). Optimal water conditions: pH 7.0-8.0, temperature 24-30°C, and conductivity of 150-600µS. These quality parameters were continually monitered by the zebrafish facitiy manager/technician and any values that fell outside of ideal ranges were corrected promptly. Zebrafish were fed twice daily with dry food (Zebrafish Management Limited, Hampshire, UK) and once daily with Artemia (Brine Shrimp Direct, Ogden, UT, USA) that had been cultured for 48 hours. Juvenile fish (up until six weeks of age) were additionally fed rotifers.

Zebrafish strain details

Two strains of transgenic reportor zebrafish were used in these experiments; Tg(bactin:GFP)(Heart reporter), and Tg(sox10:GFP)(Neural crest reporter).

Zebrafish breeding

In order to obtain embryos for these experiments, adult zebrafish were set up in breeding containers in male/female pairs the afternoon prior to egg collection as described by Westerfield1 in more detail. The two fish were kept seperated by a plastic insert until the following morning where the insert was removed and the fish allowed to breed. The inner container of the breeding system has slots in the bottom, that allows eggs to fall through into the lower compartment to avoid consumption by the adults. To collect the eggs from this lower compartment after breeding, the adult zebrafish were removed into a fresh container, and the water containing the eggs poured through a fine sieve. The eggs were then rinsed from the sieve into a petri dish containing 1x E3 and either incubated at 28°C, or used immediately for microinjection.

2.8.2 Preparation of ppa2 RNA for microinjection

To generate synthetic ppa2 mRNA, the pExpress-1-ppa2 construct (Appendix B.1) was digested with NotI and capped mRNA was synthesized in vitro by using the SP6 mMESSAGEmMACHINE Kit (Ambion) as described by Moore et al[135].

2.8.3 Preparation of PPA2 RNA for microinjection

Plasmids containing wildtype or mutant human PPA2 genes were synthesised and cloned using GeneArt (Life Technologies) into a pCS2+ vector. Construct, vector and cloning details are provided in Appendices B.2 and B.3.

2.8.4 Morpholino preparation

Two independent missense mopholino oligonucleotides (MOs) (GeneTools LLC, OR, USA) were used to transiently knock down ppa2 protein function in developing zebrafish embryos. These ∼20bp nucleic acid analogues act by binding to complementary sequences of mRNA, blocking the access of other molecules. The two MOs used in these experiments targeted the ppa2 5UTR region (including the initiation codon) to block translation, and the ppa2 exon 5-6 splice junction to modify pre-mRNA splicing. An additional morpholino was ordered as a negative control; PPA2 Splice 5mispair. This has a 5-base mismatch when compared to ppa2 Splice MO should not cause any ppa2 knockdown. Morpholino sequences are shown in Appendix B.1. The MOs arrived as lyophilised specimens that were reconstitued with Danieau buffer to give stocks of 2mM. These stocks were subsequently diluted in Danieau buffer to give working concentrations as required.

2.8.5 Microinjection of mopholino and/or RNA

Morpholinos and/or RNA were delivered into embryos that were positioned in channels of a 1.2% agarose mold (made with 1 x E3 media instead of TAE/TBE) in a petri dish. Microinjection was performed using a YOU-1 manipulator set (Narishige Co. Ltd, Tokyo, Japan). Injection needles were made using 1.0mm x 0.58mm borosilicate glass capillaries (Harvard Apparatus; Edenbridge, Kent TN86HE) pulled by a horizontal micropipette puller (Sutter Instrument Co, Novato, CA, USA). Approximately 1ng of the appropriate working concentration of MO, RNA, or MO + RNA was injected into either the yolk (MO), or the single cell (RNA and MO+RNA) of embryos at the 1-2 cell stage. Following microinjection, embryos were transferred from the agarose mold back into a petri dish containing 1 x E3 media and incubated at 28°C until the embryos had developed to the desired stage for analysis and imaging.

2.8.6 Dechorionation of embryos

Up until 48 h.p.f the zebrafish embryo is contained within an outer chorion. This may be removed manually prior to experiments on the embryo by the use of ultra fine-tipped forceps. Under magnification, one pair of foceps was used to stabilise an embryo while the other pair was used to pierce and split the outer chorion, freeing the developing embryo.

2.8.7 Characterisation of phenotype

Measurement of shortening faction (%)

Still images were captured from video clips of zebrafish hearts at maximum ventircular sytole (VS) and maximum ventricular diastole (VD). The width of the ventrical was measured in pixels using ImageJ software and the shortening faction calculated using the formula: (max VS/max VD)*100 = Shortening faction.

Heartrate measurement

Heartrate measurements were estimated from >10 second video clips of beating embryo hearts. Rate was calculated using the formula: (Number of beats counted/duration (secs))*60 = Beats per minute.

Chapter 3

Analysis of an intellectual disability pedigree

3.1 Introduction

In any case of intellectual disability (ID) which appears to only affect male individuals in a family, the possibility of an X-linked genetic cause is often the first consideration for inheritance. Female carriers in the family will normally be unaffected due to dosage compensation of their second normal X chromosome. The pedigree investigated in this chapter present just such a family - five affected males, and one unaffected female sibling (Figure 3.1). It was considered likely that the five males would all share a mutation on the X-chromosome which would account for their phenotype. As the focus of any variant analysis in this pedigree would initially be restricted to the sex chromosome, exome analysis of the proband only (rather than multiple affected family members) was undertaken. The price of sequencing and analysing one individual is significantly lower than multiple family members, however, the analysis loses the power gained from being able to reduce a candidate variant list to only variants shared by affected individuals. As this thesis explores the implementation of exome analysis (among other technologies) in a diagnostic setting, cost and time considerations are critical. This case study compares the approach of investigating one individual against the more common multi-family

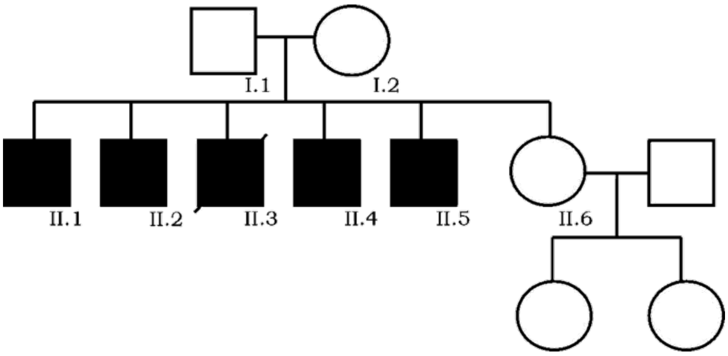


Figure 3.1 Pedigree with multiple affected individuals with ID
Moderate intellectual disability affects five males of this Caucasian family (II.1-II.5), who have estimated Wechsler Adult Intelligence Scale (WAIS) IQ scores below 50. Their female sibling (II.6) has normal intellect, and has two unaffected daughters.

member analysis.

3.1.1 Case history

A 57 year old man (patient II.2 in Figure 3.1) with intellectual disability and behavioural problems which had worsened with age was referred to a geneticist. Investigation into the family history revealed that he was one of 5 brothers who all suffered from impaired intellect. Interestingly, there is one unaffected female sibling, leading the geneticist to consider a potential X-linked genetic cause.

Patient II.2 had been previously assessed at age 42 by a psychologist and was given a mental age of 3 years 11 months. He was unable to speak clearly at times and could not read or write. He did have reasonable skills in eating, toileting, and washing, and could use the local public transport system. In addition to intellectual disability, patient II.2 was unreliable in social interactions; displaying both verbal and physical aggression at times. A chromosome microarray (arrayCGH) and Fragile X genetic testing were requested, both of which returned 'normal' results.

Based on the family history recorded by the geneticist, patient II.1 had a slightly higher functional level with a Wechsler Adult Intelligence Scale (WAIS) IQ score of 50. He had poor articulation and could not read or write, but did have some understanding of numbers and counting. He was very cooperative and social and worked well with little supervision, although requiring residential care. Patient II.3 died aged 58. Patient II.4 lives in residential care but no further information is available. Patient II.5 had been assessed to have a mental age of 4 years and 3 months. He was reported to be cooperative and attentive, but overly familiar with strangers and had some behavioural problems. The assessment made by the geneticist (A. Kidd, personal communication, 2012) was that patient II.2 and his brothers most likely have a non-syndromal X-linked intellectual disability.

3.1.2 Proband exome analysis

WES was initially only performed on a sample from patient II.2. A total of 71177 variants in 15508 genes were identified. The decision to attempt analysis on a single individual in this family was based on the assumption that this was likely an X-linked disorder. The analysis could then be restricted considerably by excluded any variants not located on the X chromosome, reducing the number of variants to 759. Splice variants in IonTorrent data sequencing were excluded from analysis as they produced a large number of false positive calls which required exclusion later. Details of analysis parameters are available in the Methods chapter (Table 2.1). Three genes were identified as harbouring potentially pathogenic mutations; *DACH2, MAGEE1*, and *TBC1D25* (Table 3.1).

Table 3.1 Candidate variant details
Locus, genotype, and reference gentoype for the three candidate variants identified in exome analysis.

Locus	Proband Genotype	Reference Allele	Type	Gene	Variant
ChrX:86067863	C	G	SNV	DACH2	c.1245G>C (p.Glu415Asp)
ChrX:75649405	C	G	SNV	MAGEE1	c.1082G>C (p.Gly361Ala)
ChrX:48417710	T	C	SNV	TBC1D25	c.729C>T (p.Gly243Gly)

The variant in *MAGEE1* was determined to be a sequence error caused by poor quality at the end of reads when the sequence read pileup was visualised in IGV, and was discarded. The variant in *TBC1D25* was confirmed in IGV, however it was actually a synonymous variant. Although this does not exclude a pathogenic effect, our ability to follow up on synonymous variants is currently limited and since *TBC1D25* has not been previously associated with intellectual disability, it was not considered a strong candidate. The *DACH2* variant was not immediately excluded in IGV visualisation or by reference sequence comparison and although there is no literature definitively implicating *DACH2* as a gene involved in intellectual disability (it is more commonly described in premature ovarian failure [136, 137], there is one reference [138] to a 325kb duplication spanning nearly the whole gene in an affected individual. *DACH2* was therefore a candidate gene further considered for investigation based on this analysis.

3.1.3 *DACH2* Analysis

DACH2 was a promising candidate gene, as a c.1245G>C (p.Glu415Asp) variant within this gene had passed filtering parameters designed to identify damaging mutations. *In silico* evaluations of pathogenicity also supported this variant as potentially pathogenic/damaging and thus a candidate for disease in this individual (Table 3.2).

Table 3.2 In silico pathogenicity predictions for DACH2:p.Glu415Asp

SIFT	PolyPhen-2	MutationTaster	PROVEAN	PhyloP	Grantham	MAF
0.07	0.412	0.979	-	0.412	45	0.001

Sanger sequencing was performed to confirm the presence of p.Glu415Asp in the proband, as well as genotype the two additional brothers for which DNA was available. Interestingly, this showed that the segregation of the variant is not consistent with the presentation of ID. Although the proband (II.2) and a second sibling (II.4) were confirmed to be hemizygous for the p.Glu415Asp variant, the third sibling (II.1) was hemizygous for the wildtype allele (Figure 3.2).

An X-linked disorder require that all affected males have inherited the same maternal X chromosome with the paternal Y chromosome. The differing genotypes of the *DACH2* variant in these siblings has two implications for X chromosome inheritance. Either *DACH2* is located in a region of the X-chromosome which has been affected by recombination, or the siblings have not inherited the same maternal X allele (Figure 3.3). Both of these scenarios exclude *DACH2* as the pathogenic cause of ID in this family, and if alternate X allele inheritance has occured then this also excludes any X-linked genetic cause and a wider analysis of autosomes should be undertaken assuming an alternative inheritance pattern. It should be noted that sample mix-up was not excluded in the non-concordant brother as a second sample was not able to be obtained.

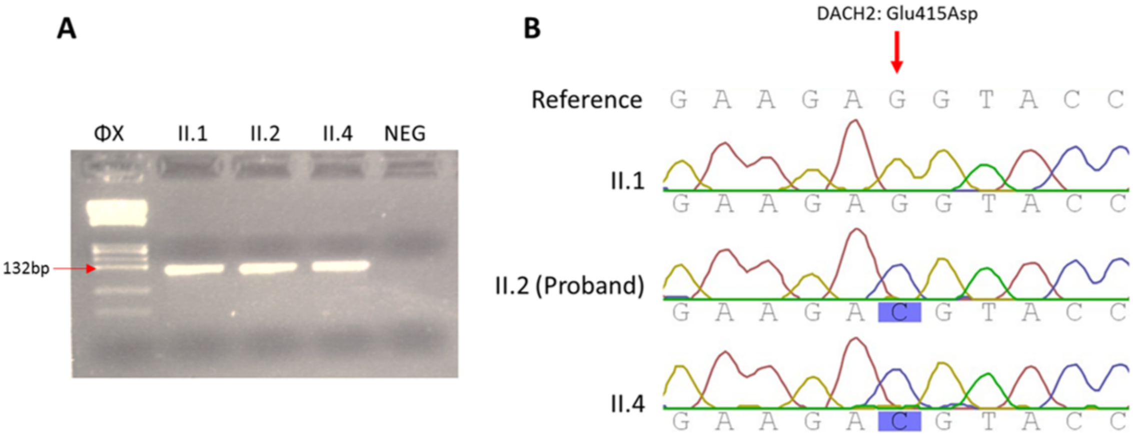


Figure 3.2 Amplification and segregation of DACH2:p.Glu415Asp
A The region of *DACH2* encompassing the residue Glu415 was amplified in individuals II.1, II.2 and II.4. The amplification products were visualised on a 2% agarose gel against a molecular weight standard produced from a HaeIII digest of the ϕX 174 bacteriophage. **B** Sanger sequencing of these products showed that the p.Glu415Asp variant does not segregate with disease in this pedigree.

3.2 Multi-sample exome analysis

Following the results generated by the single proband approach described above in the previous section, it was decided that additional exome analysis on two additional brothers (for whom DNA was available) would be a valuable extension of the analysis thereby providing enough

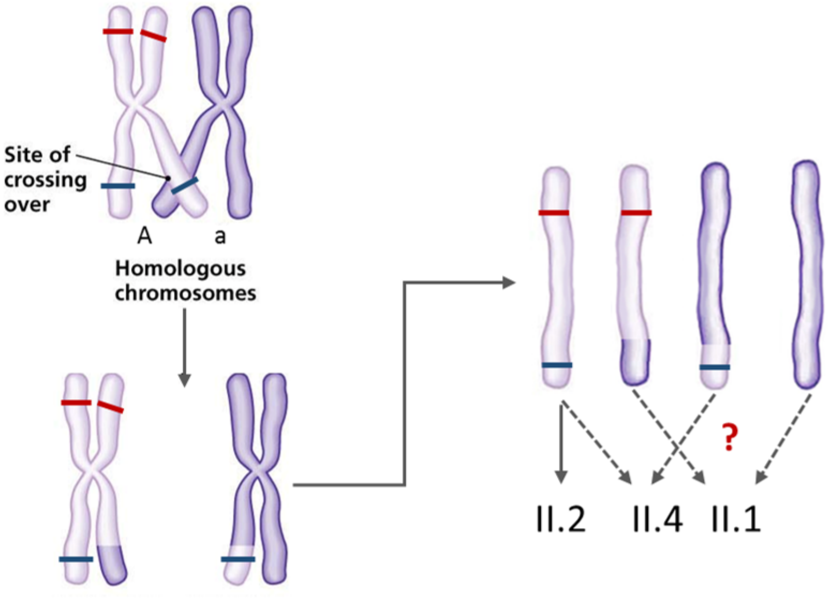


Figure 3.3 Mechanism of recombination of the X chromosome
Recombination occurs between homologous chromosomes prior to gamete formation. The red bar represents a pathogenic mutation on one maternal allele, while the blue bar represents another SNP (in this case *DACH2*:p.Glu415Asp) that is analysed. The fact that individual II.1 does not share the p.Glu415Asp variant as individuals II.2 and II.4 indicates that **either** recombination has occurred between this and the causative loci, **or** that the siblings have inherited alternate X alleles and the cause of ID in this pedigree is not X-linked. This diagram also illustrates how siblings II.2 and II.4 may not have inherited the same X allele despite sharing the *DACH2* mutation.

power to examine variants in the autosomal chromosomes. For this analysis a combination of in-house scripts and the Ionreporter software was employed to identify candidate variants in parallel.

Based on the broader autosomal mechanism of disease now proposed, the autosome analysis was initially targeted to a panel of known ID genes (Appendix C.1). A perl script was written by Zoe Ward to identify variants in genes on these panels from the exome data. This was initially run on only the proband, resulting in 123 variants (SNVs and indels) which was narrowed down to three genes where compound heterozygosity of variants may be occurring; MAPT (Microtubule-associated protein tau), SLX4 (SLX4 structure-specific endonuclease subunit) and KMT2D (Lysine (K)-Specific Methyltransferase 2D).

A second script was used to identify variants shared by all three brothers. This list contained

6564 variants that were interrogated against the ID panel, though this produced no instances of predicted compound heterozygosity of likely pathogenic variants. The same script was also used to identify variants in known autism-associated genes (Appendix C.2), included due to the phenotype overlap and produced 4 candidate genes; *NCKAP5* (NCK-associated protein 5), *KMT2C* (Lysine (K)-Specific Methyltransferase 2C), *HYDIN* (HYDIN, axonemal central pair apparatus protein) and *MYH4* (myosin, heavy chain 4, skeletal muscle).

Parallel analysis in the IonReporter software had also produced a list of 66 variants, which included 5 candidate genes (containing two or more variants shared between all three brothers) under loose filtering conditions; *HYDIN, KMT2C, KRT6A* (keratin 6A, type II), *ZNF717* (zinc finger protein 717) and *FRG2C* (FSHD region gene 2 family, member C). Due to its reoccurrence under different filtering conditions *KMT2C* was identified as the strongest candidate gene for further investigation in this family, specifically the variants c.871C>T (P.Leu291Phe) and c.2512G>A (p.Gly838Ser). Other less highly ranked variants were catalogued for potential follow up analysis, beyond the scope of this thesis.

The possibility of sample mix-up that had been highlighted in the X-chromosome analysis of *DACH2* was deemed unlikely after analysis of these additional exomes, as Y chromosome variants between the three brothers were 100% concordant.

3.2.1 *KMT2C* Analysis

The presence of two *KMT2C* variants identified by both the IonReporter based analysis, and the comparison to 'candidate' gene panels, meant that this was considered a promising candidate for further analysis. *In silico* predictions of pathogenicity for both variants were variable, but several methods suggested that the amino acid change would be pathogenic (Table 3.3).

Sanger sequencing was performed to confirm that both the p.Leu291Phe and p.Gly838Ser variants detected in WES data in the three brothers (II1,II.2 and II.4), and additionally in the forth

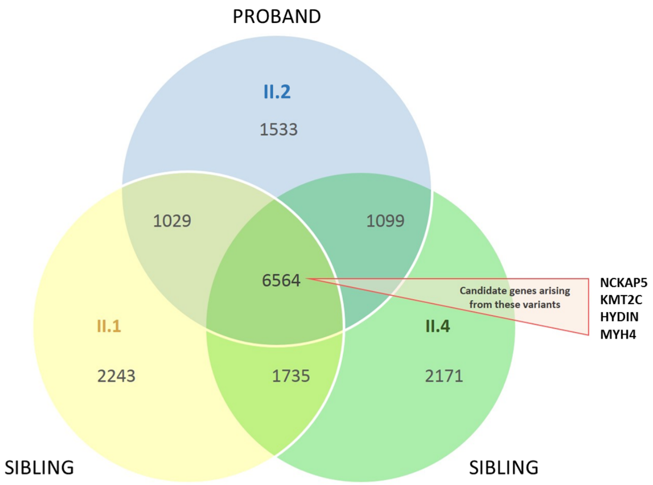


Figure 3.4 Sibling exome variant filtering - Autism and ID panel script written for this thesis
A script was written to mimic the IonReporter function to display and filter variants that were common to multiple individual analyses. The list of 6564 variants shared between individuals II.1, II.2 and II.4 was then interrogated based on the pre-defined ID and Autism panels (See Appendix D) and filtered to exclude genes that contained less than two variants. No variants were identified in genes listed on the ID panel, but four Autism panel genes harboured shared variants in all siblings; *NCKAP5, KMT2C, HYDIN and MYH4.*

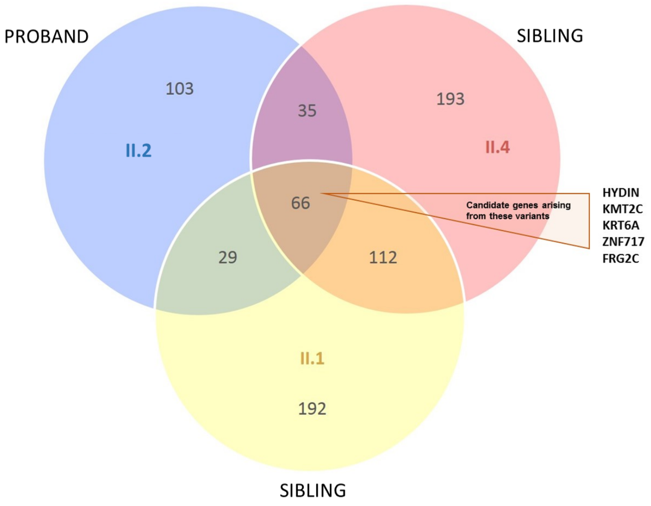


Figure 3.5 Sibling exome variant filtering - IonReporter
Filtering for possible causal recessive variants within these siblings was performed using the Venn diagram function within IonReporter that allows identification of overlapping variants between analyses on seperate individuals, in this case II.1, II.2 and II.4. When the filter chain described in the Methods (Table 2.2) was run on these samples, the overlapping number of variants was 66. Between 103-193 additional variants were only identified in one sibling. A further 176 variants were shared by 2/3 of the siblings.

Table 3.3 In silico prediction of pathogenicity for KMT2C
Pathogenicity prediction scores from six *in silico* software and minor allele frequency (MAF) as defined by the ExAC (http://exac.broadinstitute.org/).

KMT2C	SIFT	PolyPhen-2	MutationTaster	PROVEAN	PhyloP	Grantham	MAF
p.Gly838Ser	0.0	1	1	-	2.66	56	$1.65E^{-5}$
p.Leu291Phe	0.05	1	1	-	2.43	22	-

Damaging (Pink); Moderate (Orange); Neutral (Green)

surviving brother (II.5) where DNA was acquired after exome analysis was performed. Initial sequencing results for both variants did not support the exome sequencing results. The variant p.Leu291Phe appeared homozygous rather than heterozygous (Figure 3.6), and the variant p.Gly838Ser appeared absent entirely (Figure 3.7) in all four brothers.

To examine whether these results may have been generated by non-specific PCR amplification or other causes of allele amplification failure that we have previously investigated ([139], the PrimerBLAST tool was used to identify unintended targets that may be amplified with the primers used across the p.Gly838Ser variant. Along with the expected result for *KMT2C*, there was an additional complementary region identified within chromosome 21 (Figure 3.8). Alignment of the nucleotide sequence between the expected *KMT2C* region (NG_033948.1) and the chromosome 21 region (NC_000021.9 (10473535-10499510) from the GRCH38 p2 primary assembly) indicated strong conservation. The original primers used to sequence these two *KMT2C* variants did not differentiate between the intended chr 7 target and the homologous region on chr 21. Allele specific primers were designed to cover regions of difference between the two sequences (Figures 3.9) and specifically amplify *KMT2C*. *KMT2C*-specific resequencing of both variants identified in the exome analysis confirmed that neither variant was present in these brothers, but were instead an artifact of the chromosome 21 *BAGE* gene family reads mis-aligning to *KMT2C* during the mapping process (Figure 3.11).

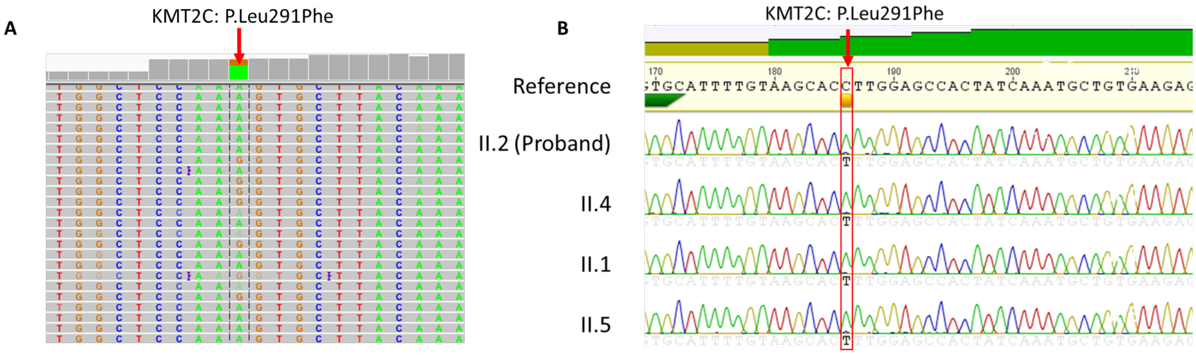


Figure 3.6 Sequencing of KMT2C; p.Leu291Phe
A IGV visualisation of exome sequencing read pileup from individual II.2 indicating heterozygousity for a C>T nucleotide change resulting in p.Leu291Phe. **B** Sanger sequencing electropherogram of the same individual indicating homozygosity of the T allele. Note the reverse orientation of display between the IGV output and the electropherogram.

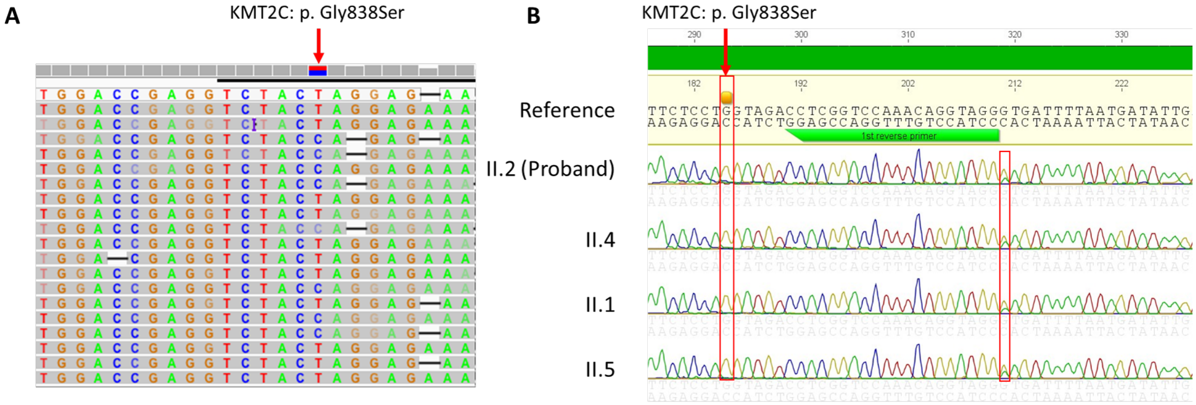
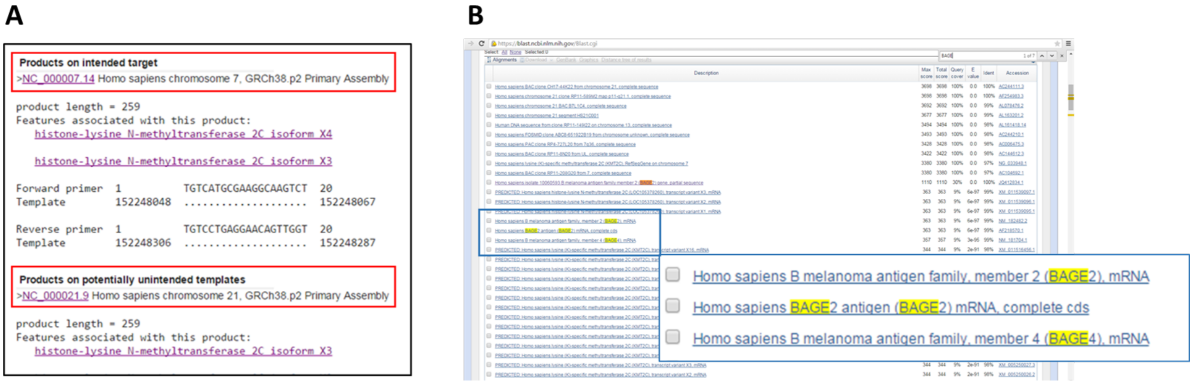


Figure 3.7 Sequencing of KMT2C ; p.Gly838Ser
A IGV visualisation of exome sequencing read pileup from individual II.2 indicating heterozygosity for a T>C nucleotide change resulting in p.Gly838Ser. **B** Sanger sequencing electropherogram of the same individual indicating no nucleotide change. Heterozygosity at a second allele in the electropherogram is also non-concordant with the IGV visualisation, prompting investigation. Note the reverse orientation of display between the IGV output and the electropherogram.



Products on intended target
>NC_000007.14 Homo sapiens chromosome 7, GRCh38.p2 Primary Assembly

```
product length = 259
Features associated with this product:
   histone-lysine N-methyltransferase 2C isoform X4

   histone-lysine N-methyltransferase 2C isoform X3

Forward primer  1          TGTCATGCGAAGGCAAGTCT  20
Template        152248048  ....................  152248067

Reverse primer  1          TGTCCTGAGGAACAGTTGGT  20
Template        152248306  ....................  152248287
```

Products on potentially unintended templates
>NC_000021.9 Homo sapiens chromosome 21, GRCh38.p2 Primary Assembly

```
product length = 259
Features associated with this product:
   histone-lysine N-methyltransferase 2C isoform X3
```

☐ Homo sapiens B melanoma antigen family, member 2 (BAGE2), mRNA
☐ Homo sapiens BAGE2 antigen (BAGE2) mRNA, complete cds
☐ Homo sapiens B melanoma antigen family, member 4 (BAGE4), mRNA

Figure 3.8 Chromosome 21 homologous region (BAGE)
A PrimerBLAST was used to identify potential unintended amplification targets which might explain the abberant sequencing results (see Figures 3.6 and 3.7), a region on chromosome 21 was identified with a histone-lysine N-methyltransferase feature which is also a possible target of the primers used to amplify chromosome 7 *KMT2C* variants. **B** The chromosome 21 region aligning with *KMT2C* showed hits for B melanoma antigen (BAGE) gene family members with 99%-100% identity. This corresponds with the known history of BAGE genes, that they arose from a duplication of *KMT2C*.

3.3 Discussion

This chapter was intended to investigate the power of exome analysis in one affected individual, when the causal mutation was likely to be X- linked, rather than a trio or multi-sibling analysis. However, after initial analysis indicated that the underlying genetic cause could not be identified on the X- chromosome the project was extended to further family members and an analysis of all chromosomes (WES).

Originally, analysis of the proband's exome identified one candidate X-chromosome gene, *DACH2*, harbouring a missense mutation p.Glu415Asp. Multiple *in silico* predictions for this variant supported a pathogenic effect, as did the rare MAF score. However, further investigation of this variant by Sanger sequencing in two additional affected brothers did not support *DACH2* as the causative gene, as only two of the three siblings that were sequenced, shared the variant. No DNA was available on the two additional brothers. This also indicated the possibility that not all of the 5 brothers had inherited the same maternal X-chromosome, and that an X-linked disease may not be the inheritance pattern after all.

Two further brothers underwent exome sequencing, and the analysis of the three siblings was

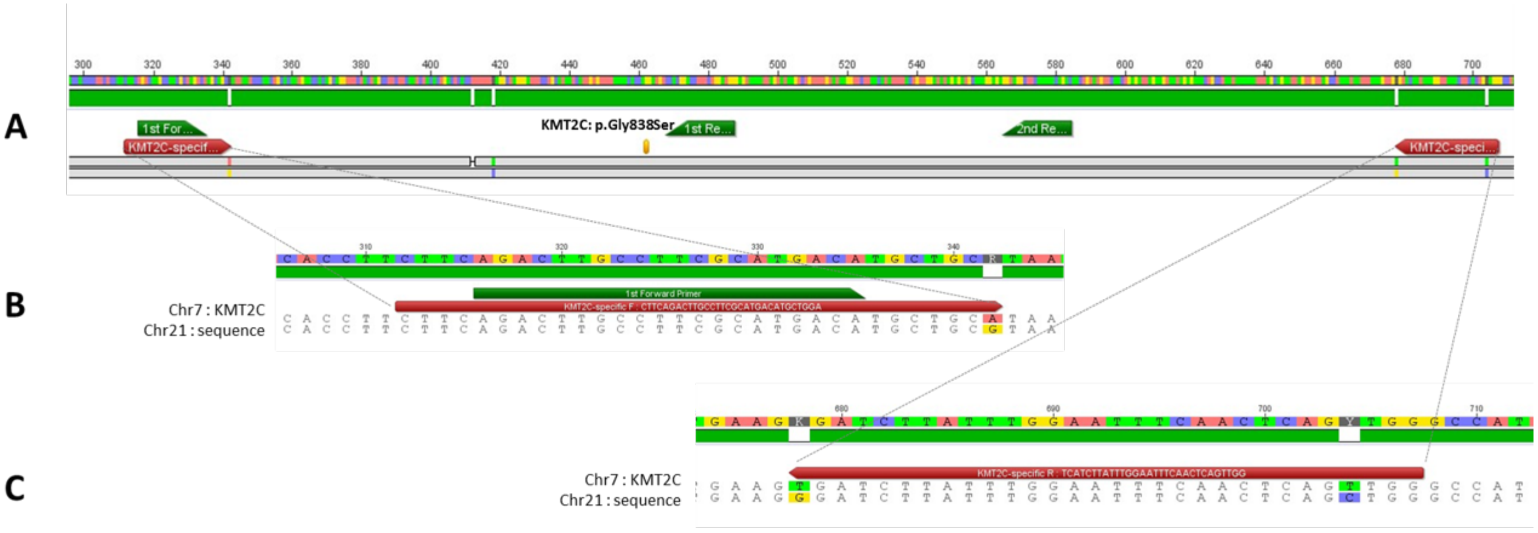

A
B
C
300
320
340
360
380
400
420
440
460
480
500
520
540
560
580
600
620
640
660
680
700
1st For...
KMT2C-specif...
KMT2C: p.Gly838Ser
1st Re...
2nd Re...
KMT2C-speci...
310
320
330
340
1st Forward Primer
KMT2C-specific F : CTTCAGACTTGCCTTCGCATGACATGCTGGA
Chr7 : KMT2C
Chr21 : sequence
CACCTTCTTCAGACTTGCCTTCGCATGACATGCTGCATAA
CACCTTCTTCAGACTTGCCTTCGCATGACATGCTGCGTAA
680
690
700
710
KMT2C-specific R : TCATCTTATTTGGAATTTCAACTCAGTTGG
Chr7 : KMT2C
Chr21 : sequence
GAAGTGATCTTATTTGGAATTTCAACTCAGTTGGGCCAT
GAAGGGATCTTATTTGGAATTTCAACTCAGCTGGGCCAT

Figure 3.9 KMT2C; p.Gly838Ser allele specific primer design

A Alignment of p.Gly838 showing nucleotide conservation between *KMT2C* and homologous chromosome 21 region. **B** Positioning of the allele specific Forward primer to terminate at a nucleotide variation between *KMT2C* and the homologous chromosome 21 region.**C**Positioning of the allele specific Reverse primer to terminate at a nucleotide variation between *KMT2C* and the homologous chromosome 21 region, also incorporating a second nucleotide variation for added specificity. Note both allele specific primers were also designed with mismatched penultimate bases at the 3' end.

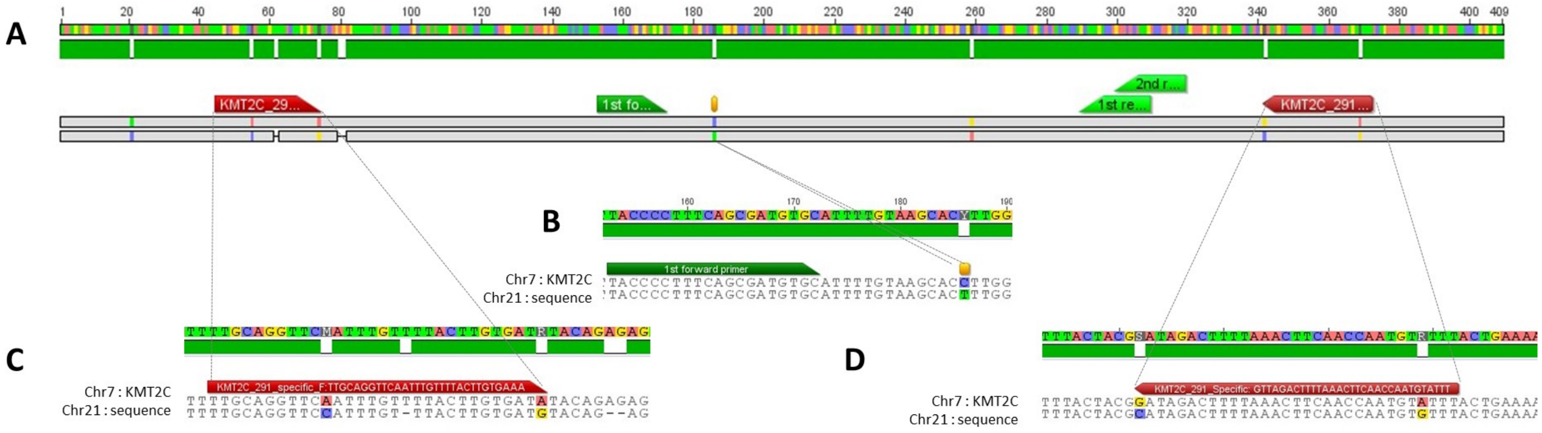

A
1
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
409
KMT2C_29...
1st fo...
2nd r...
1st re...
KMT2C_291...
B
160
170
180
190
1st forward primer
Chr7 : KMT2C
Chr21 : sequence
TACCCCTTTCAGCGATGTGCATTTTGTAAGCACCTTGG
TACCCCTTTCAGCGATGTGCATTTTGTAAGCACTTTGG
C
KMT2C_291_specific_F:TTGCAGGTTCAATTTGTTTTACTTGTGAAA
Chr7 : KMT2C
Chr21 : sequence
TTTTGCAGGTTCAATTTGTTTTACTTGTGATAATACAGAGAG
TTTTGCAGGTTCCATTTGT-TTACTTGTGATGTACAG--AG
D
KMT2C_291_Specific: GTTAGACTTTTAAACTTCAACCAATGTATTT
Chr7 : KMT2C
Chr21 : sequence
TTTACTACGGATAGACTTTTAAACTTCAACCAATGTATTTACTGAAA
TTTACTACGCATAGACTTTTAAACTTCAACCAATGTGTTTACTGAAA

Figure 3.10 KMT2C; p.Leu291Phe allele specific primer design

A The positioning of original non-specific primers (green) and the allele specific primer position (red) are indicated. **B** Yellow highlighting indicates nucleotide 2512 is a variable position between the aligned sequences. The c.871C >T(P.Leu291Phe) variant identified in the exome data may conceivably be a alignment error if chromosome 21 reads mapped to *KMT2C*. **C** Positioning of the allele specific Reverse primer to terminate at a nucleotide variation between *KMT2C* and the homologous chromosome 21 region, also incorporating a second nucleotide variation for added specificity. Note the allele specific primer was also designed with a mismatched penultimate base at the 5' end.

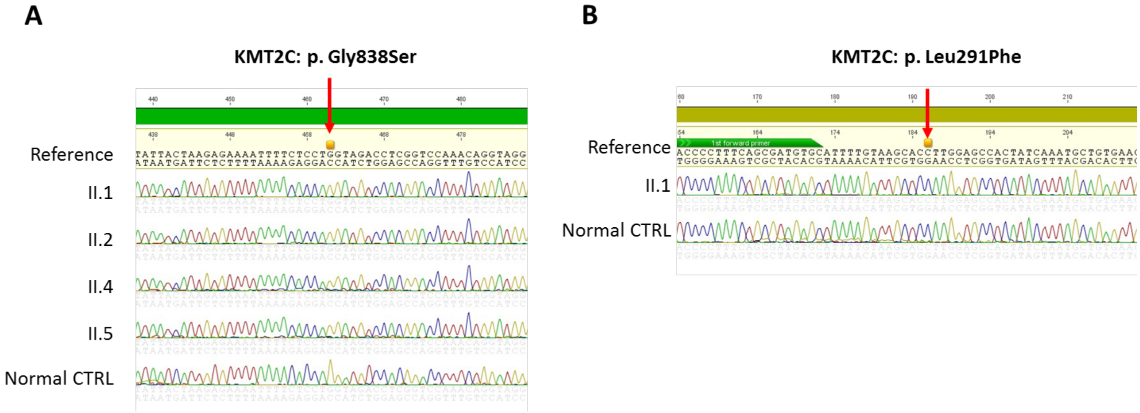


Figure 3.11 KMT2C- specific Sanger sequence
A Specific sequencing for KMT2C residue 838 **B** Specific sequencing for KMT2C reside 291

performed independently by two methods. The first involved using the ThermoFisher Scientific software IonReporter, and the second used in-house Perl scripts to mimic the analysis processes that IonReporter was carrying out automatically, but with a focus on specific gene lists relating to autism and ID. The scope of these analyses was extended from the X-chromosome to all chromosomes. Both analysis approaches identified the gene KMT2C as a candidate gene for further investigation. This gene contained two variants (p.Leu291Phe and p.Gly838Ser) shared by all three brothers analysed, raising the possibility of a recessive model of inheritance if the variants were compound heterozygous. Both variants were predicted by various *in silico* models to be probably pathogenic, and further investigation and confirmation with Sanger sequencing was performed on all three siblings.

Interestingly, the Sanger sequencing did not correlate with the exome sequencing data, and neither mutation could be confirmed as heterozygous. The earlier variant (p.Leu291Phe) appeared to be homozygous for the mutant allele in all three brothers and an unrelated healthy control, while the second variant (p.Gly838Ser) was not present in any individual when Sanger sequencing was undertaken. This discrepancy indicated an issue with the specificity of the PCR, and it was noted that the primers used to amplify *KMT2C* were also targeting a homologous region on chromosome 21. This region relates to a member of the B melanoma antigen (BAGE) gene family which originally arose from a duplication of *KMT2C* [140, 141], explaining the

90-100% identity seen between these sequences.

KMT2C-specific primers were designed to ensure that no amplification of *BAGE* was occurring. *KMT2C*-specific resequencing of both variants confirmed that neither variant was truly present in these brothers, but were instead an artefact of the chromosome 21 *BAGE* gene family reads mis-aligning to *KMT2C* during the mapping process of the exome data processing.

No other significant candidate gene mutations were identified in this family which would account for the ID phenotype in all 5 brothers, although identifying candidates largely relied on prior knowledge of ID-associated genes. This approach is not a good way to identify novel genes, and would have missed a diagnosis such as the one presented further on this thesis (Chapter 5). Another limitation of this analysis is the difficulty of investigating mutations present on the Y-chromosome. Approximately 78 protein coding genes have been identified on the Y-chromosome, and the function of many is still unknown [142]. There are however, several genes that are expressed mainly in the brain such as *PCDH11Y* and *NLGN4Y* which are involved in formation of dendrites and axons in the CNS [143]. Some Y-chromosome rearrangement, duplication and deletion events have been associated with ID [144, 145], but specific genes are rarely implicated. Exome coverage of this chromosome is hampered by the highly repetitive nature of the male specific region (MSY) of the Y-chromosome, and the X-homologous regions at the telomeric tips.

If this family is harbouring a molecular defect in the Y-chromosome, then the exome analysis performed here will likely not be able to detect this event.

Another explanation for not detecting a causative variant, is that it occurs outside the coding regions sequenced in the current analysis. MPS has recently been successfully employed to identify a non-coding regulatory mutation which is associated with moderate X-linked ID in a Northern European family [146]. In this case, the causative variant had been mapped to Xq28-qter, and a custom capture array was designed to target all coding regions and putative functional non-coding regions in the linkage interval identifying a mutation in the intergenic region between *HCFC1* and *TMEM187* which was not visible on exome sequencing. Due to the massively increased data produced by whole genome sequencing (WGS), this was not considered a feasible or appropriate use of resources in the case of this family and due to the absence of parental samples or more distant relatives, a linkage approach by other methods would be more difficult. The five patients are in their 5th or 6th decade, and a diagnosis is unlikely to significantly affect their management. The unaffected parents are deceased, and therefore there is no clinical imperative to identify a causative mutation for pre natal screening purposes. There would be a considerable benefit for the unaffected sister in this family, given that she has produced two healthy female offspring who may go on to have children at-risk of ID but there is little interest from these family members in the analysis performed here, and DNA would likely not be obtained.

In the context of a diagnostic laboratory, this case would currently fall beyond the scope of feasibility, as neither the time or resources would be available to investigate this case to the extent it might otherwise deserve. Until WGS is a practical alternative that can be routinely performed, cases like this may remain undiagnosed unless new approaches or candidate genes become available to change the prediction weightings and modify our analysis to produce new findings which may have been discarded by the algorithms used for this analysis.

Chapter 4

Trio analysis - a sporadic case of premature ageing

4.1 Introduction

Exome analysis of parent-child trios is a technique which has previously identified casual mutations in neurodevelopmental cases such as autism spectrum disorder (ASD) and intellectual disability (ID) [147, 133, 148]. Reported success rates for a diagnostic result are significantly improved among families undergoing trio sequencing as opposed to a single individual analysis, 37% compared to 21% respectively according to one US clinical laboratory [133]. For patients with sporadic disease a diagnosis can be particularly challenging and trio sequencing provides more information on inheritance patterns of variants than singleton sequencing. This chapter investigates whether trio sequencing of an affected proband and her unaffected parents could detect an underlying mutation causative of a premature ageing phenotype and provide a unifying diagnosis.

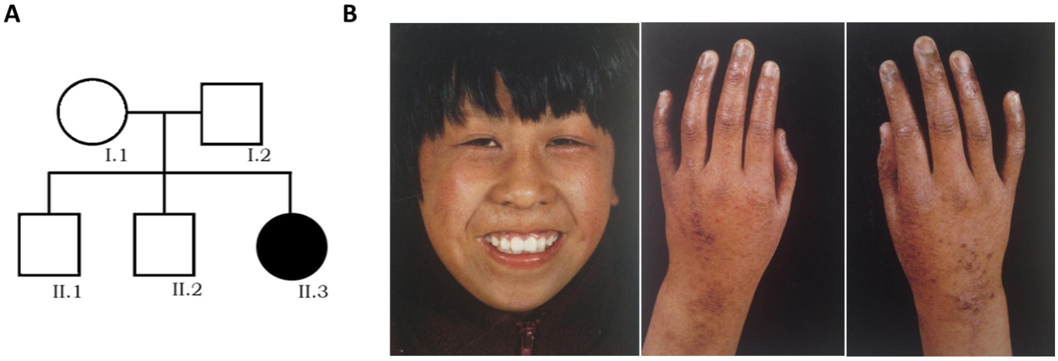


Figure 4.1 Pedigree and morphology
A The proband is the third child born to healthy non-consanguineous Chinese parents. The proband suffers from a syndrome that includes global developmental delay and macrocytosis among other symptoms. She has two unaffected brothers. **B** Unusual facies, nails, and skin abnormalities of the proband.

4.2 Case history

An 8 year old girl (individual II.3 in Figure 4.1) presented with an ageing syndrome including primary adrenal failure, microcephaly with global developmental delay, hyperactivity, behavioural problems, short stature, dyserythropoietic macrocytosis (Hb 118g/L), recurrent otitis media and several skin conditions including reticulate pigmentation, keratotic papular skin lesions and atopic eczema. Notable facial features included short palpebral fissures, a smooth filtrum and a flattened nasal bridge. The description of an 'aging disorder' related to her brittle nails, evidence of greying hair, and general skin appearance. Her parents are both of Chinese ancestry and were not known to be related. She has two brothers, neither of whom are affected.

Cytogenetic analysis using G-banding showed a normal female karyotype 46XX, and an arrayCGH analysis was also normal. Fanconi's anemia (FA) was considered but excluded as no excessive chromosome breakage was observed in peripheral blood cells cultured with mitoycin C (a specific feature of FA). A diagnosis of dyskeratosis congenita was considered, but determined to be unlikely as there was no evidence of the characteristic shortening of telomeres seen in this disorder. Exome analysis was considered an appropriate step to take in this case where a unifying diagnosis could potentially improve the treatment and management of the patient, and have implications for other family members.

4.3 Exome analysis

Exome sequencing of the family trio was performed on an Ion Proton instrument. Analysis of the variants was performed in the cloud-based software IonReporter (https://ionreporter.thermofisher.com/ir/), utilising the inbuilt Trio exome analysis pipeline which distinguishes the proband from the mother and father, allowing identification of compounding recessive or rare *de novo* mutations predicted to have deleterious impact.

4.3.1 Recessive analysis

To investigate a possible recessive inheritance pattern of disease, an analysis was performed to identify genes in which compounding mutations are present in the proband. Details of analysis parameters are available in the Methods chapter (Table 2.3). This analysis identified only one candidate gene, MIR4273 (microRNA 4273). Two trans-phase MIR4273 variants were present in the proband while both parents were heterozygotes (Table 4.1). Unfortunately, assigning function to this microRNA through the available literature was not possible (beyond an association with risk of colorectal cancer [149]), and no further analysis was performed on this variant.

Table 4.1 Candidate variant - recessive

Locus	Proband Genotype	Father Genotype	Mother Genotype	Reference Allele	Type	Gene
Chr3:75786802	G/A	G/A	G/G	G	SNV	MIR4273
Chr3:75787996	C/T	C/T	C/C	C	SNV	MIR4273

4.3.2 *De novo* analysis

A second approach to the analysis was to identify possible pathogenic variants occurring in a *de novo* manner in the proband (therefore not observed in either parent). Details of analysis parameters for this approach are available in the Methods chapter (Table 2.4). This identified candidate variants in three distinct genes considered to be relevant to the observed phenotype; *ZNF595* (Zinc Finger Protein 595), *PRIM2* (Primase, DNA, Polypeptide 2), and *BCOR* (BCL6 corepressor) (Table 4.2).

ZNF595 encodes a protein belonging to the Cys2His2 zinc finger protein family, whose members function as transcription factors that can regulate a broad variety of developmental and cellular processes [150]. *PRIM2* encodes the 58 kilodalton subunit of DNA primase, an enzyme that plays a key role in the replication of DNA. The subunit functions as a DNA-directed RNA polymerase to synthesize small RNA primers that are used to create Okazaki fragments on the lagging strand of DNA [151]. The protein encoded by *BCOR* was identified as an interacting corepressor of BCL6, a POZ/zinc finger transcription repressor that is required for germinal centre formation and may influence apoptosis [152]. Diseases associated with *BCOR* include oculofaciocardiodental syndrome (OFCD), myelodysplastic disorders, and Lenz microphthalmia syndrome (LM) [153]. Visualisation of the read pileup across these six candidate variants in IGV showed that all but the heterozygous Gly1303Ser variant in *BCOR* were false positive calls by the software. This was attributed to either poor coverage of the variant site, highly repetitive regions, or the variant being inherited but not called in the carrier parent due to either of the two prior reasons.

Table 4.2 Candidate variants - sporadic

Locus	Proband Genotype	Father Genotype	Mother Genotype	Reference Allele	Gene	Variant
Chr4:59350	A/G	A/A	A/A	A	ZNF595	N/A
Chr6:57398226	A/G	T/G	T/G	T	PRIM2	N/A
ChrX:39922265	C/T	C/C	C/C	C	BCOR	c.3907G>A (p.Gly1303Ser)

False positive variants in *ZNF595* and *PRIM2* were excluded at this point. The *BCOR* variant, c.3907G>A (p.Gly1303Ser), was confirmed as present in the proband but neither parent in IGV (Figure 4.2).

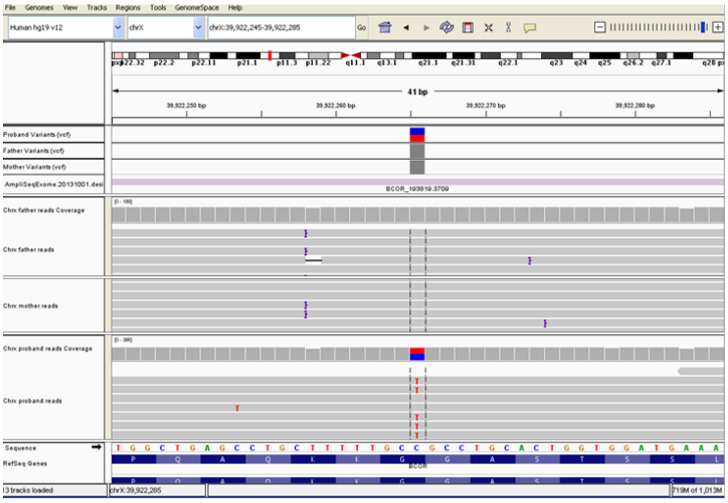


Figure 4.2 BCOR; p.Gly1303Ser - IGV visualisation
This variant was confirmed to be heterozygous in the proband when the exome read pileup was visualised in IGV. There was no evidence of the variant in either parent.

4.4 Investigation of *BCOR*, BCL6 co-repressor

4.4.1 Sanger confirmation

Sanger sequencing was employed to confirm the presence of the *BCOR* variant in the proband, and also the *de novo* classification by examining the parents for absence of the variant in the germ-line. This approach confirmed that the proband was heterozygous for c.3907G>A (p.Gly1303Ser) in DNA extracted from both blood and buccal cell, while both parents were homozygous for the wild type G nucleotide (Figure 4.3).

4.4.2 Allele-specific Sanger sequencing

Because Sanger sequencing is not sensitive enough to reliably distinguish somatic from germline mutations, apparent *de novo* mutations may in fact have arisen during gametogenesis in one of the parents, or very early in the proband's embryonic development. In order to investigate the possibility of parental mosaicism, a nested allele-specific PCR assay was designed to preferentially amplify either the wild type or mutant allele. Initial attempts to amplify proband *BCOR* using these highly specific primers failed, and optimisation of annealing temperature did not resolve the issue. Since non-allele-specific amplification of the region had been achieved using a

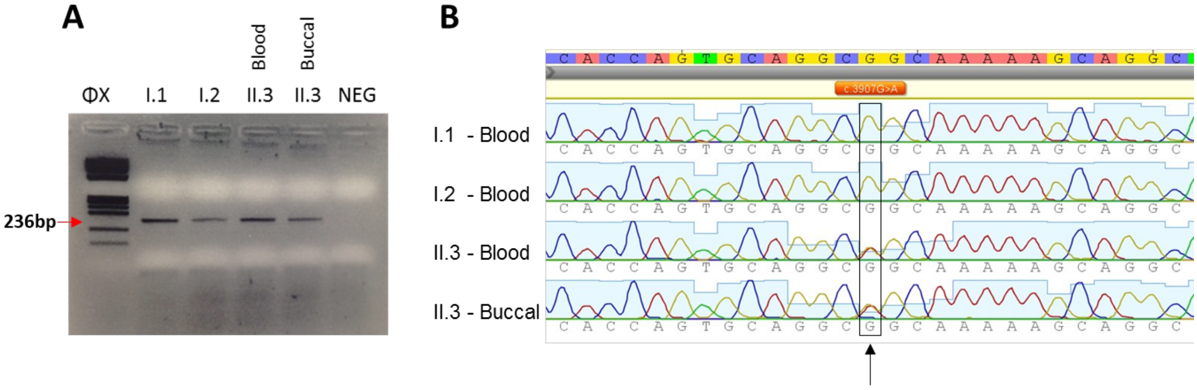


Figure 4.3 BCOR (Gly1303Ser) confirmation
A Agarose gel showing PCR products amplified from the parents of the proband, and blood and buccal DNA from the proband prior to sequencing of the products. **B** Sanger sequencing electropherogram of the amplified products. The variant c.3907G>A (p.Gly1303Ser) is present in both blood and buccal DNA in the proband (II.3) but neither parent (I.1 and I.2). No evidence of mosaicism in the proband is evident based on the two tissues analysed.

previous primer set (BCOR F1 and BCOR R1) (Figure 4.3), a combination of BCOR R1 and the allele specific forward primers (WASP MUT and WASP WT) were employed with proband DNA as the template, although this also resulted in failed amplification (data not shown).

A nested approach was utilised, using a dilution of the amplicon produced with BCOR F1 and BCOR R1 (PCR product 1) as the template material for a second amplification with the allele-specific primers. This nested PCR was performed over a temperature gradient of 55°C to 65°C (Figure 4.4). Amplification of the WT-specific product remains clear in all three family members even at 65°C. Importantly, background amplification in the two parents drops out at the higher annealing temperature, while no change is observed in the proband. An annealing temperature of 65°C was used in all subsequent PCR to ensure reduction of any non-specific amplification.

A further nested PCR was performed using two dilutions of this product; 10^{-4} (Figure 4.5A) and 10^{-6} (Figure 4.5B) to assess sensitivity of the PCR to detect mutant product. At both dilutions, there is clear amplification of WT-specific bands in all individuals, and clear amplification of mutation-specific product in the proband with no evidence of mutant product in the unaffected samples (Figure 4.5).

Table 4.3 In silico prediction of pathogenicity for BCOR; p.Gly1303Ser

SIFT	PolyPhen-2	MutationTaster	PROVEAN	PhyloP	Grantham	MAF
0	0.804	0.89	0.481	0.43	56	$5.24E^{-5}$

Damaging (Pink); Moderate (Orange); Neutral (Green)

As a final measure to exclude the presence of any mutant amplification in either parental blood, the nested PCR using PCR product 1 template dilution 10^{-6} was repeated in an extended 50-cycle protocol. The number of PCR cycles in this protocol should amplify even very small amounts of mutant allele present in the diluted template. No amplification of mutant product was observed in either parent or normal control, although some non-specific amplification was observed at a lower molecular weight indicating possible primer dimerisation or off taget effects due to a high number of cycles. The proband displayed a strong mutant product at 118bp. A strong wildtype product was observed in all four samples, and absent in the negative control with no evidence of non-specific amplification or primer dimerisation (Figure 4.6). This does not definitively rule out parental gonadal mosaicism as this method has limited sensitivity, particularly when using DNA extracted from peripheral blood.

4.4.3 Frequency and mutation effect

In silico tools were used to predict the pathogencity of the *BCOR* variant c.3907G>A (p.Gly1303Ser). SIFT and MutationTaster both predicted the variant to be damaging to the protein's function, with scores of 0 and 0.89 respectively. PROVEAN predicted it was a neutral change, while the PolyPhen, PhyloP and Grantham scores predict moderate damage. At the time of investigation, the ExAC Browser MAF for this variant was 0.00005236 (1 allele in 190970)(Table 4.3). All of these points of evidence indicate that further investigation into this variant was warranted.

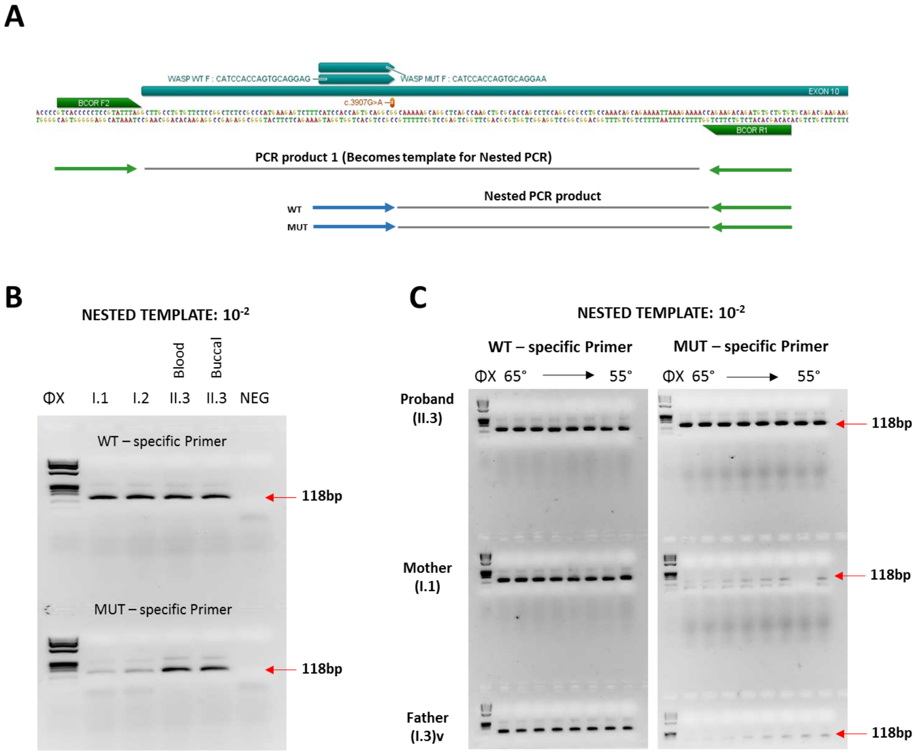


Figure 4.4 BCOR allele-specific nested PCR optimisation
A Schematic of the primers used to product the initial product, and the nested PCR product which is then visualised on the gels below. PCR product 1 was amplified using the original non-specific primers, then used as the template for a nested PCR with the WT/MUT allele specific primer specifically amplifying either the wild type or mutant allele if present. **B** Lane 1 - Proband (II.3), lane 2 - Mother (I.1), lane 3 - Father (I.2), lane 4 - Normal control, lane 5 - Negative control. PCR product 1 at a 10^{-2} dilution was used as the template material in a nested PCR (60° annealing) with either the wild type or mutant allele-specific primers with a common non-specific reverse primer. Some non-specific amplification is still visible in both parental samples. **C**. This nested PCR was repeated over a temperature range from 55°C to 65°C. Amplification of the WT-specific primer, and the MUT-specific primer (in the proband) were observed even up to 65°C, and at this temperature there was significantly reduced non-specific bands present in the parental samples. As the annealing temperature approaches 55°C increased background appears on the gel, as seen in B.

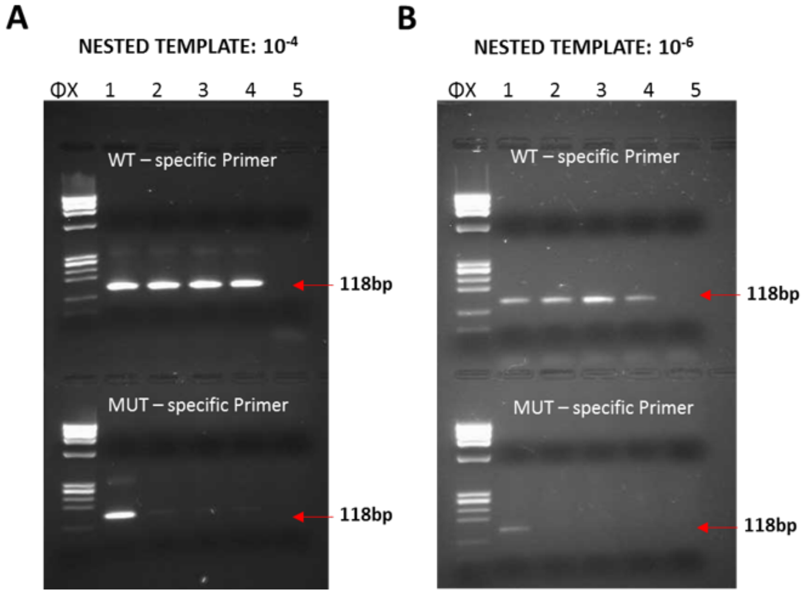


Figure 4.5 BCOR nested PCR-template dilutions
Lane 1 - Proband (II.3), lane 2 - Mother (I.1), lane 3 - Father (I.2), lane 4 - Normal control, lane 5 - Negative control. **A** Nested PCR (65° annealing/31 cycles) using PCR product 1 diluted 10^{-4} as template material, shows clear amplification of WT-specific bands and a MUT-specific band in the proband. The two parents and the unrelated normal control show no amplification or significant background at this dilution/annealing temperature. **C** The specificity of the nested PCR was further examined by using PCR product 1 diluted 10^{-6} as the template material. Even at this dilution, there is clear amplification of WT-specific bands in all individuals, and clear amplification of MUT-specific product in the proband with no evidence of mutant product in the unaffected samples.

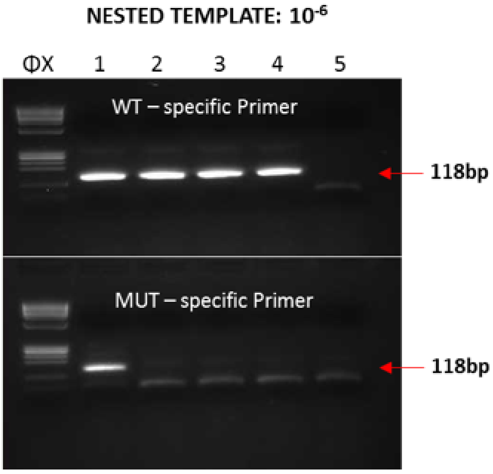


Figure 4.6 BCOR allele specific PCR optimisation
Lane 1 - Proband (II.3), lane 2 - Mother (I.1), lane 3 - Father (I.2), lane 4 - Normal control, lane 5 - Negative control. Nested PCR (65° annealing/ 50 cycles) using PCR product 1 diluted 10^{-6} as the template material. Extended cycling of these products shows clear amplification of WT-specific bands in all individuals, but only the proband shows any amplification of MUT-specific product. Note that gel has been cropped at point indicated with white line.

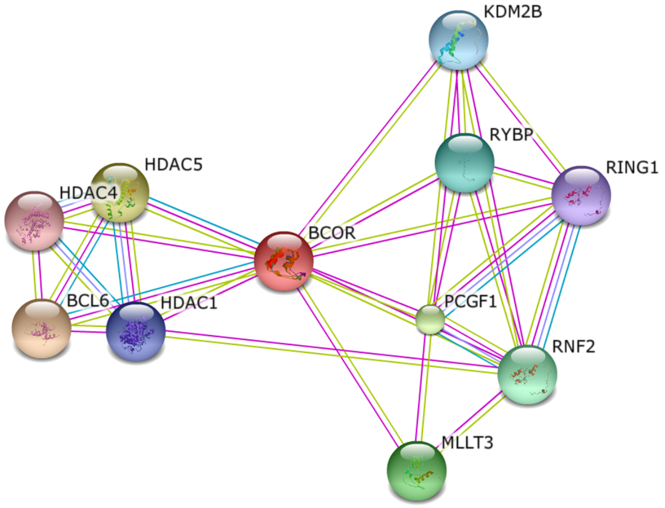


Figure 4.7 BCOR associations
Interactions of *BCOR* generated by STRING (Known and predicted Protein-Protein Interactions v.9.1: http://string91.embl.de) evidence view. Interacting proteins of *BCOR* include the BCL6 complex proteins RING1, PCGF1, KDM2B, RYBP, BCL6 and HDACs as well as MLLT3 (Myeloid/Lymphoid Or Mixed-Lineage Leukemia Translocated To, 3).

4.4.4 *BCOR* functional role

BCOR encodes an interacting corepressor of the BCL6 complex (Figure 4.7). The complex also includes the proteins ring finger1B (RING1B), polycomb group ring finger1 (PGCF1) and KDM2B, a lysine-specific demethylase. RING1B catalyses the monoubiquitylation of H2AK119, while the binding of PCGF1 and KDM2B stimulates the E3 ligase activity of RING1B. KDM2B recruits the BCL6 complex to non-methylated CpG islands, where it functions to remove histone H3K36me2 and induce repressive histone H2A monoubiquitylation, enhancing BCL6-mediated transcriptional repression [154] (Figure 4.8).

The disruption or dysregulation of proper transcription control is recognised as a major cause of cancers and developmental disorders [155, 156]. *BCOR* appears to be ubiquitously expressed in all tissues, based on supporting evidence from literature [152] and experimental data collection from the EMBL-EBI Expression Atlas (a subset of data shown in Figure 4.9).

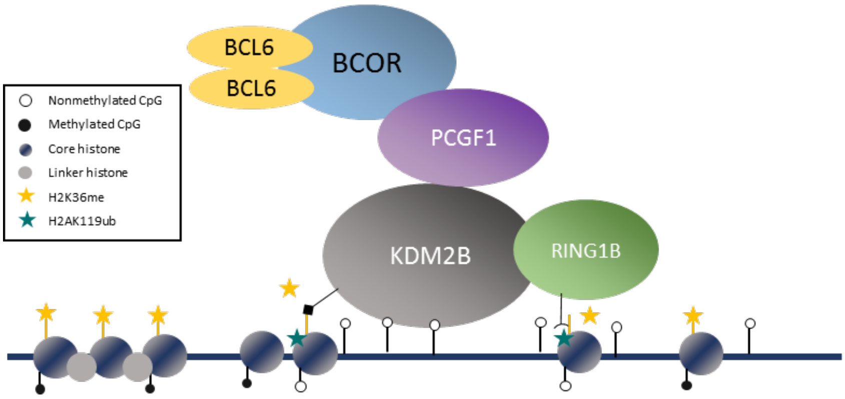


Figure 4.8 The BCL6 complex
The BCL6 complex acts as a unique repressive, transcriptional machinery at CpG islands within transcription start sites. Nonmethylated CpG islands are used to recruit KDM2B which then recruits the BCL6 complex resulting in H3K36me2 depletion and repression of histone H2A monoubiquitylation in these regions.

4.4.5 *BCOR* disease association

BCOR mutations have been associated with human disease, although not previously with the phenotype presenting in individual II.3. Oculofaciocardiodental syndrome (OFCD) is a rare X-linked dominant condition with male lethality, characterized by microphthalmia, congenital cataracts, facial dysmorphic features, congenital heart defects, and dental anomalies [157]. Ng *et al* were first to identify and report frameshift, deletion and nonsense mutations in *BCOR* in seven OFCD syndrome pedigrees. They also reported one missense *BCOR* mutation (p.Phe85Leu) in two male patients with Lenz microphthalmia (LM) syndrome (characterized by abnormal development of the eyes and several other parts of the body) [158, 159]. These findings support the role of *BCOR* as a key transcriptional regulator during early embryogenesis [160], as well as regulating adult mesenchymal stem cell (MSC) functions via an epigenetic mechanism [161].

To date, OFCD and LM are solely caused by mutations in *BCOR* (although *BCOR* mutation are not considered the major molecular cause of LM). All *BCOR* mutations implicated in OFCD so far are null alleles that introduce premature stop codons and transcripts are presumed to be degraded by nonsense mediated decay. These mutations are shown along with the *de novo* variant found in this study in Figure 4.10. A study by Hilton *et al* indicated that OFCD is a spectrum disorder, and at the mild end may have relatively insignificant non-ocular features

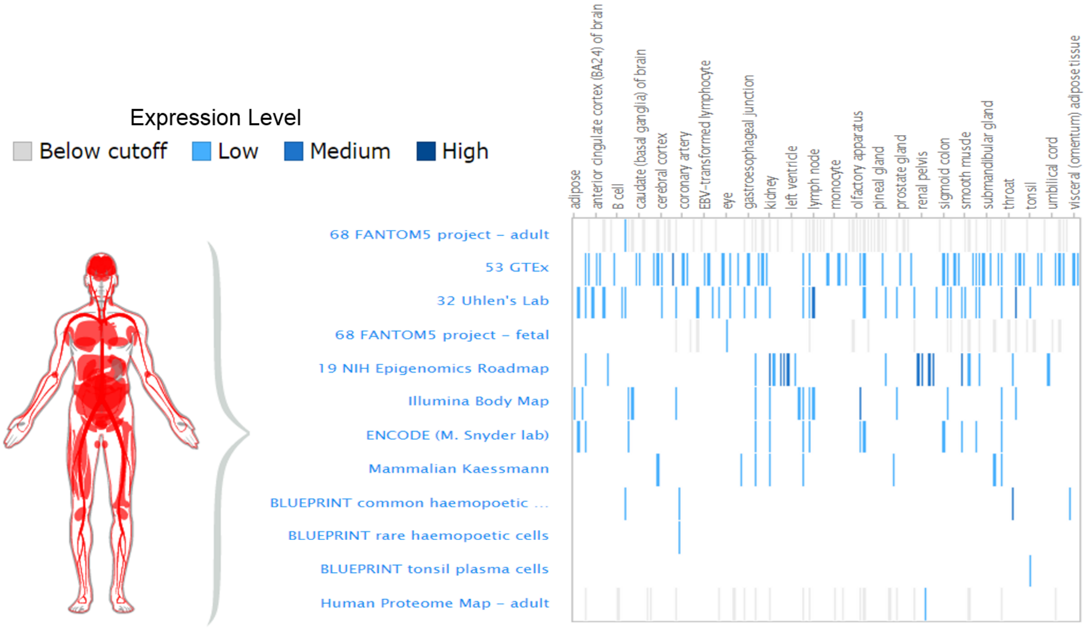


Figure 4.9 BCOR expression in tissues
Visual extract of from the EMBL-EBI Expression Atlas (http://www.ebi.ac.uk/) showing a summary of the expression data cataloged from twelve experimental sources, providing evidence for ubiquitous expression of *BCOR* in human tissues (ID: ENSG00000183337).

[159]. Truncated BCOR protein is not detected in OFCD patients, likely due to nonsense-mediated mRNA decay of the transcript [158]. OFCD is therefore considered a condition of haploinsufficiency rather than a dominant negative effect of mutant *BCOR* in sensitive cells [162, 163, 164].

Chimeric fusion transcripts of *BCOR* have been found in a broad range of cancers, including a fusion transcript with the retinoic acid receptor alpha (RARA) in acute promyelocytic leukemia (APL). Somatic mutations in *BCOR* have been detected in acute myeloid leukemia (AML) patients with normal karyotypes (CN-AML), myelodysplastic syndrome (MDS) and chronic myelomonocytic leukemia (CMML) patients [154]. Somatic *BCOR* mutations have been associated with poor prognosis in MDS and CN-AML patients [165, 166]. Additionally, somatic *BCOR* mutations have been identified in 3.3% of patients with medulloblastoma [167] and 9.5% of patients with retinoblastoma [168]. These findings imply that *BCOR* is a tumour suppressor gene, and that the BCOR complex signalling may be an important theraputic target in *BCOR* complex-dysregulated cancers, as well as showing that the BCOR complex is necessary for proper cell differentiation and proliferation [154].

4.5 X-inactivation

During early human female development, one X chromosome is inactivated in each cell to compensate for the difference in X chromosome gene dosage between males and females. This is typically a random process resulting in approximately half of the cells expressing the maternal or paternal X allele. Non-random X-inactivation has been observed in some females, including symptomatic carriers of an X-linked recessive disease and asymptomatic carriers of an X-linked dominant disorder [169]. In female patients with OFCD, investigation of peripheral blood lymphocytes show a strong bias towards inactivation of the mutant *BCOR* allele-carrying X chromosome. This was seen in six individuals with OFCD in whom X-inactivation studies were informative, all showing skewing of 90-100% towards the non-mutant *BCOR* allele, indicating that *BCOR* is essential for normal hematopoiesis [158]. The varying degree of skewed X-inactivation which can occur between tissue types, and even between OFCD patients would explain some of the heterogeneity of phenotype seen in OFCD.

To investigate X-inactivation patterns in this pedigree, the highly polymorphic trinucleotide repeat within the first exon of the human androgen receptor gene (AR) and the methylation-sensitive restriction enzyme *Hpa*II were used to distinguish between the maternal and paternal X alleles and simultaneously determine their methylation. The X-inactivation assay was performed on samples from the proband's mother (peripheral blood DNA), the proband (peripheral blood DNA and buccal cell DNA) as well as a known non-random X-inactivation control to investigate potential skewed inactivation in this pedigree (Figure 4.11). In a healthy female individual with random X-inactivation the ratio of each allele ranges from 50:50 to 70:30. Moderate skewing is usually defined as a ratio above 80:20, while extreme skewing is a ratio above 90:10 [169]. In this analysis the peripheral blood DNA sample of the proband showed moderate skewing

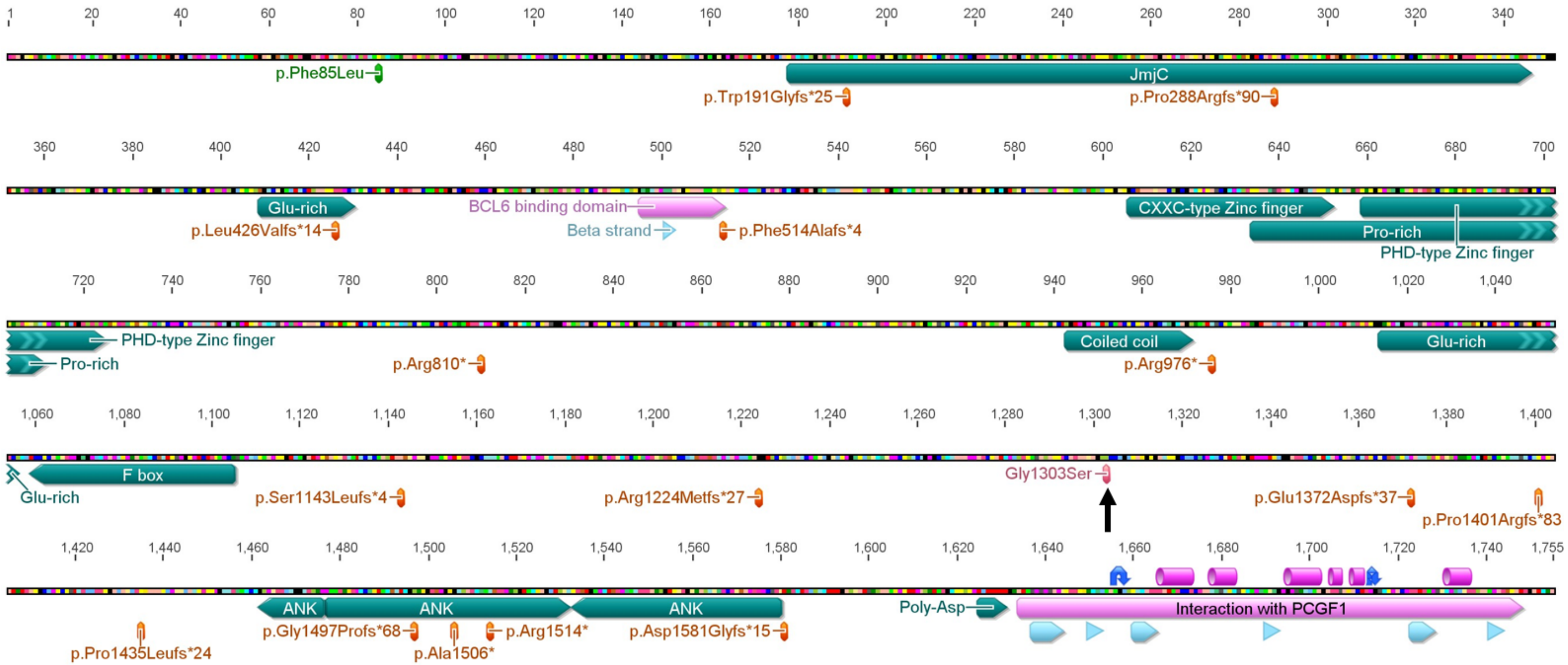


Figure 4.10 BCOR protein annotation
The *de novo* variant (p.Gly1303Ser) identified in this study (indicated by arrow) as well as previously documented pathogenic mutations [159] in *BCOR* (OFCD-associated in orange, LM-associated in green) are annotated alongside protein features and domains. JmjC:Jumonji C domain, ANK: Ankyrin repeat domain, Blue arrow: turn, Fuchsia tube;helix.

just below the cutoff for extreme skewing (13:87), while her buccal DNA showed random X-inactivation (42:58) similar to that of the healthy, age matched female control (41:59). The proband's mother showed moderate skewing with a ratio of 19:81. The frequency of moderate skewing increases with age [170], and although the proband and her mother both show similar levels of skewing, this is likely to be more significant in the proband who is only in her second decade of life.

Differentiation between the maternal and paternal *BCOR* allele in the proband revealed that the paternal allele was preferentially inactivated (maternal allele is reduced after digestion with *Hpa*II which targets the active allele) . Since the *BCOR* variant appears to be *de novo* in the proband, the parental allele cannot be determined. Therefore, we can only predict that skewed inactivation of the paternal X chromosome is due to presence of the *BCOR* variant. This is consistent with evidence that a father confers significantly more *de novo* mutations to their children than do mothers [171].

4.6 FA exclusion

The finding of skewed X-inactivation, preferentially inactivating the paternal X chromosome, prompted further consideration of a previous diagnostic assay performed on this individual to exclude Fanconi's anemia (FA). FA is the most common type of inherited bone-marrow failure syndrome with an incidence of 1 per 350,000 births [172]. FA results from mutations in a number of genes involved in DNA repair mechanisms. These mutations increase the cell's sensitivity to cross-linking agents (such a mitomycin C), and diagnosis is often made on observing chromosome breakage after treatment with one of these agents [173, 174]. The clinical features of FA are varied, including growth retardation, malformations of the kidneys, heart and skeleton, microcephaly, hypogonadism, cutaneous abnormalities (hyper/hypopigmentation, cafe-au-lait spots) and unusual facies. Life threatening symptoms are bone marrow failure (BMF) and a predisposition to cancer. There are currently 17 genes known to cause FA. The majority of these genes are responsible for DNA repair at stalled replication forks generated either spontaneously

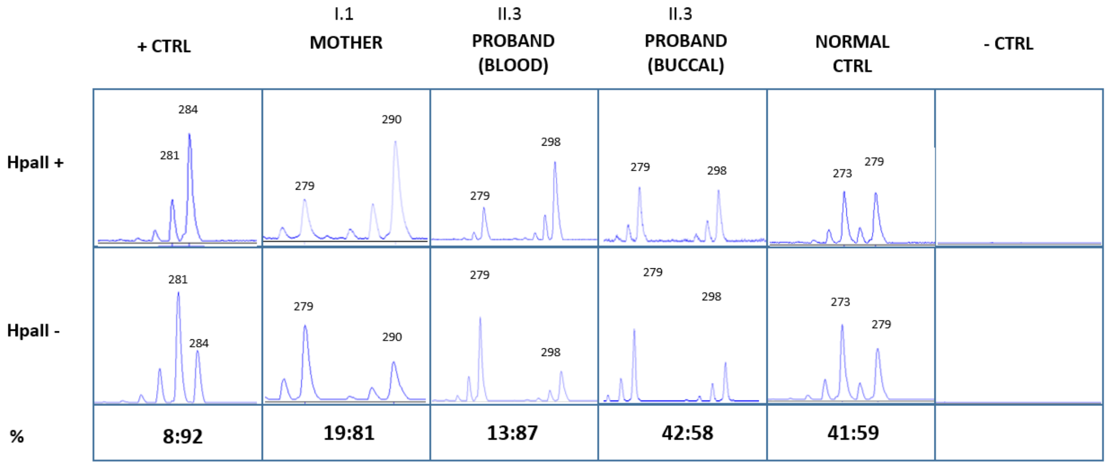


Figure 4.11 X-inactivation assay for non-random skewing
Analysis of the relative inactivation ratio of the X alleles in this pedigree based on the amplification of the highly polymorphic (CAG)$_n$ repeat within the 5' end of the coding region of the human androgen receptor gene (AR), revealed moderate/significant skewing (13:87) in peripheral blood DNA from the proband, while DNA from a buccal sample appeared to have random X-inactivation (42:58). Sizing of the alleles between the maternal sample and the proband indicates that inactivation of the paternal allele is occurring in the proband's peripheral blood sample. The proband's mother also displays moderate skewing with an allele ratio of 19:81. Near complete skewing (8:92) is visible in the positive control sample, previously determined to have skewed X-inactivation. A healthy female individual was included as a normal control, and random X-inactivation is visible as an allele ratio of 41:59. X allele ratios are reported as averages of four replicates.

or by DNA interstrand cross-links and other DNA damage [172].

A positive result for FA has been previously reported in one female with OFCD [175] and the authors of this report raised the question of *BCOR* mutations potentially being implicated in FA, particularly as a possible causes of AML development often seen in these patients [176]. When FA analysis was initially performed in this patient, prior to exome sequencing, peripheral blood cells were cultured in the presence of mitomycin C which induces chromosome breakage in individuals with FA. The issue arising from the discovery of skewed X-inactivation is that the percentage of cells expressing a mutant allele may be too low for detection. In individual II.3, peripheral blood cells may be expressing only approximately 10% mutant *BCOR* and 90% normal *BCOR*. Any chromosome breakage analysis performed on these cells will be biased towards the breakage observed in cells expressing a normal allele (assuming *BCOR* variant inactivation), resulting in a possible false negative result.

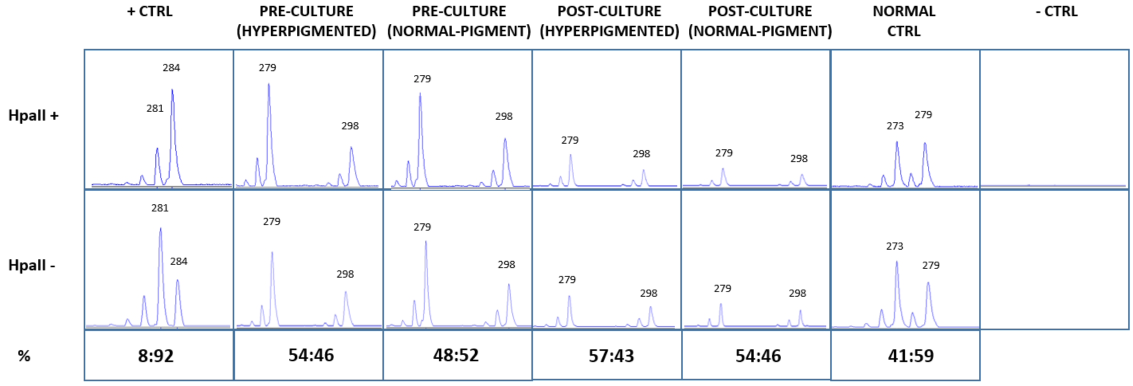


Figure 4.12 X-inactivation assay of skin biopsy
DNA was extracted from both the hyperpigmented and normal pigmentation biopsy. X-inactivation analysis was performed on both samples pre and post cell culture. No evidence of skewing was observed either pre or post culture with allele ratios <60:40. X allele ratios are reported as averages of all replicates.

In order to exclude FA as a diagnosis for this individual it was decided that the mitomycin C assay should be repeated on cultured cells which were not affected by skewed X-inactivation. A sample of the proband's buccal cells showed random X-inactivation when analysed, but a sample of buccal cells is inappropriate for culture due to the non-sterile environment of the mouth. Instead, two skin biopsies were performed by a dermatologist to obtain fibroblasts for culture. One biopsy was collected from skin with normal pigmentation, and the second from a hyperpigmented area in case X-inactivation was associated with the pigmentation phenotype. X-inactivation analysis was performed on DNA extracted from the biopsy samples immediately post collection, and then following culturing to firstly determine if skewed X-inactivation was present in either sample and secondly to ensure that no significant skewing was introduced during the culture process. No skewing was observed in either skin biopsy pre or post culturing (Figure 4.12). Both cell cultures were then sent to Cincinnati Children's Hospital who could perform a diagnostic Mitomycin C assay. Neither sample showed significant levels of chromosome breakage and were reported negative for FA.

4.7 Discussion

A mutation in *BCOR* (p.Gly1303Ser) was identified in a female patient with a syndrome characterised by premature ageing, behavioural problems, developmental delay and dyserythropoietic macrocytosis among several other symptoms. Although not entirely novel, the *BCOR* mutation is only documented once in a person of South Asian descent in the ExAC database and has not been functionally characterised. The proband has two brothers who are healthy aside from some slight learning difficulties. Prior to exome sequencing, both dyskeratosis congenita and Fanconi's anemia (FA) had been considered possible diagnoses, although they were subsequently excluded (or considered unlikely) following gold standard diagnostic testing.

Analysis of the parents indicated that this mutation is a *de novo* occurrence in the proband. *In silico* prediction scores of pathogenicity varied but largely supported a deleterious effect of the Gly1303Ser change. SIFT and MutationTaster gave scores of 0 and 0.89 respectively, indicating the mutation is likely to be damaging or disease causing. PolyPhen-2 scored the variant at 0.804 which equates to a possibly damaging prediction. The mutation frequency as assessed by ExAC was extremely rare, seen in only 1/190970 alleles, and the wildtype Gly amino acid at position 1303 is strongly conserved between species (PhyloP, 4.7; PhastCons, 1). Structural modelling of the variant was limited as there is no 3D structure of the whole gene as yet.

In the past, the field of Mendelian medical genetics has largely been restricted to inherited variation by family-based linkage studies. The more recent advances of microarrays and MPS technologies have allowed the study of *de novo* mutations, which were largely missed in previous investigative models. It is only now emerging, the extent to which *de novo* mutations play a role in both common and rare forms of neurodevelopmental disease [177]. On average, an individual's genome contains 74 germ line de novo SNVs [178], which although rare, have not been subjected to stringent evolutionary selection [179, 180] and are therefore strong candidates for sporadic disease [177]. In addition to SNVs, *de novo* indels and CNVs are present, although these occur at a lower frequency. The rate of indel occurrence equates to approximately three novel changes per genome per generation [181], and the frequency of CNVs even lower, occurring only once in every 50 individual genomes [182].

For an apparent *de novo* mutation that occurs in an affected child, it is often important for the parents future reproductive choices to ascertain whether the mutation occurred exclusively in the proband in either a germline or mosaic form, or whether it has been transmitted via parental mosaicism. *De novo* mutations can occur at any stage of gametogenesis and at further stages of development. If the affected proband is mosaic, then it is likely that the mutation arose postzygotically and the chances of recurrence in another sibling is negligible. If, however, a parent was an unaffected somatic mosaic carrier of the mutation the recurrence rate in a future child may be up to 50%, requiring very different counselling situations [183, 184]. In this pedigree, an allele specific PCR assay was designed to try and exclude the possibility that either parent is a significant somatic mosaic for the *BCOR* mutation. However, no detection of the mutation in peripheral blood does not exclude the possibility that one parent is a gonadal mosaic. Analysis of this scenario would require ovarian or sperm samples, which were considered too invasive to procure. The basis for investigating peripheral blood in the parents was considered a good compromise based on a recent study which found that apparent high quality *de novo* mutations in a proband actually arose from low level parental somatic mosaicism (in 4/50 trios) [185], or gonadal mosaicism which is detectable in the parental peripheral blood by sequencing [186]. It is important to note that deep sequencing across the *BCOR* mutation was not performed in either parent, and therefore the question of pre or post zygotic inheritance is not fully resolved in this pedigree.

Although there is a reasonable association between the *de novo* BCOR mutation identified in this proband, and the phenotype displayed, the process of establishing pathogenicity and causality of an uncharacterised variant in this context is achieved by building evidence from multiple sources. The avenues of information that can be used in establishing pathogenicity are summarised in Figure 4.13. Gene function is an important aspect to be considered, and consideration of the effect of the mutated gene in the context of the phenotypic characteristics.

The former relies on known information available, while the latter may be assessed by *in silico* approaches although with the current low level of genomic data available, *in silico* models are problematic for *de novo* missense mutations in particular. Evidence for the pathogenicity of a mutation may be built from factors such as the gene in question known to cause the specific disease, or in a related pathway. Or less strongly, mRNA expression patterns in disease affected tissues and inferred functional attributes. Data from model organisms and protein-protein interaction studies can further support this evidence [177]. An integrated approach of all lines of evidence, and the phenotype of the patient currently remains the most accurate predication tool for variant pathogenicity.

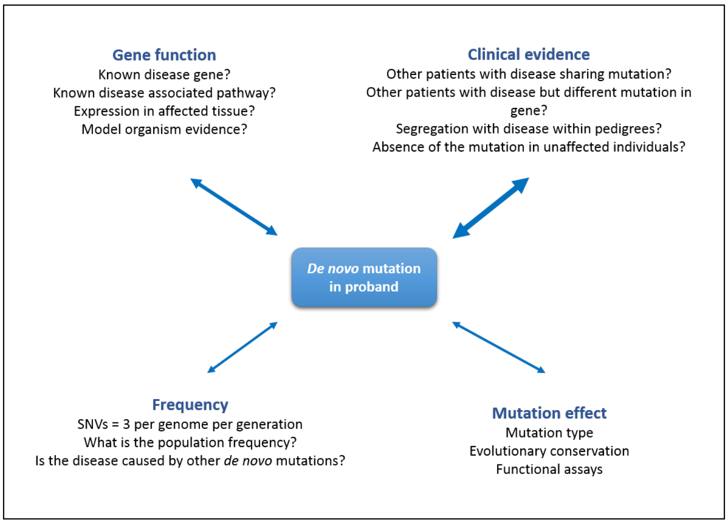

Gene function
Known disease gene?
Known disease associated pathway?
Expression in affected tissue?
Model organism evidence?
Clinical evidence
Other patients with disease sharing mutation?
Other patients with disease but different mutation in gene?
Segregation with disease within pedigrees?
Absence of the mutation in unaffected individuals?
De novo mutation in proband
Frequency
SNVs = 3 per genome per generation
What is the population frequency?
Is the disease caused by other de novo mutations?
Mutation effect
Mutation type
Evolutionary conservation
Functional assays

Figure 4.13 Establishing de novo mutation causality
One of the biggest challenges in interpreting exome data is understanding which mutations may be causal in the phenotype at hand. In the case of *de novo* mutations this can be a difficult task as irrelevant *de novo* variants are likely to outnumber disease causing mutations. There are several factors to consider when assigning disease causality to a *de novo* mutation. **Mutation effect** and **Gene function** should be considered in terms of the phenotype. Nonsense mutations in a relevant gene for example is more likely to be causal of disease than a synonymous SNV in an gene not expressed in affected tissues. **Frequency** of the mutation can also be an indication of pathogenicity, as common SNVs are unlikely to cause disease. Most compellingly would be **Clinical evidence** which would include multiple unrelated affected individuals with the same *de novo* mutation, or segregation of the mutation within a large pedigree. Figure adapted from Veltman *et al* 2012 [177].

BCOR is an important corepressor, potentiating transcriptional repression of the proto-oncoprotein BCL-6 and suppressing the transcriptional activity of AF9, a common mixed-lineage leukemia fusion partner. *BCOR* mutations have been implicated in Lenz microphthalmia (LM), oculofaciocardiodental syndrome (OFCD) and myelodysplastic disorders, and recent studies suggest that *BCOR* uses a unique combination of epigenetic modifications to direct gene silencing [187]. Characterisation of protein expression in mice shows that *Bcor* displays widespread but specific expression during embryonic development. Strong expression of *Bcor* was observed in the developing eye, tooth primordial, limb buds, branchial arches, and multiple nervous system tissue which correlates with the tissues most commonly affected in OFCD and LM patients [160, 188]. Hilton *et al* [189] additionally confirmed a role of *BCOR* in vertebrate laterality determination (defective in some cases of OFCD causing dextrocardia and intestinal malrotation) and ocular development, modeling OFCD in *Xenopus tropicalis.*

BCL-6 is a subclass of zinc finger proteins with a N terminus POZ/BTB domain and C terminus zinc fingers. BCL-6 interacts with several corepressors via these domains, with the POZ domain interacting with NCOR, SMRT and BCOR in a mutually exclusive fashion, suggesting that the functions of BCL-6 may be segregated among the corepressors. The BCOR complex and the mono-ubiquitylated form of histone H2A are shown to localise to several BCL-6 targets, including *p53 (TP53)* and *Cyclin D2 (CCND2)* in lymphoma cells, suggesting that the BCOR complex is regulating these genes through H2A monoubiquitylation activity [187]. Cyclin D2 is one of three cyclins that act as critical governors of the mammalian cell-cycle clock during the G1 phase. Cylin D2-deficient female mice are sterile as ovarian granulosa cells cannot proliferate normally in response to follicle stimulating hormone (FSH), and deficient male mice display hypoplastic testes [190]. TP53 is crucial in multicellular organisms where it functions as a tumour suppressor. In human cancer, *TP53* is the most frequently mutated gene, functionally involved in gene transcription, DNA synthesis and repair, genomic plasticity and programmed cell death. Additionally, tumours with *TP53* mutations are often more aggressive and gene expression has been implicated as an independent prognostic factor in carcinomas [191].

BCL-6 plays a critical role in specific immunological processes, acting as a transcriptional switch that controls germinal centre formation. BCL-6 deficient mice do not form germinal centres, have defects in T-cell-dependent antibody response and go on to develop inflammatory disease. Growth retardation, and prominent myocarditis and pulmonary vasculitis were observed in BCL-6 deficient mice who often died prior to 5 weeks of age [192]. Further investigation of BCL-6 null mice showed systemic inflammatory disease involving multiple organs, most commonly the myocardium, spleen, gut, liver and skin. Prominent eosinophilic infiltrates and an increase in IgG1 and IgE-bearing B cells were observed in these mice, consistent with a T helper 2 (Th2)-mediated inflammatory response [193]. Th2 cells are involved in mediating the activation and maintenance of the antibody-mediated immune response by producing cytokines which have a variety of effects on numerous cell types. Among these effects is the stimulation and recruitment of specialised subsets of immune cells in response to allergens or infection. Additionally Th2 cells stimulate the maturation of B cells into IgE-producing plasma cells. A summary of the symptoms and features observed in the proband of the affected family in this case study is provided in Section 4.2. Of interest in relation to the immunological involvement of *BCOR* are the numerous symptoms that may have an inflammatory basis such as otitis media, atopic eczema, reticulate pigmentation, general ageing appearance of skin, nails and hair, and even her primary adrenal failure. Although the p.Gly1303Ser mutation in *BCOR* found in the proband does not lie within the region of the gene that encodes the BCL-6 binding region, without a structural model it is uncertain whether there may still be a deleterious effect of the mutation on BCL-6 binding when the protein folds. It is therefore possible that this mutation in *BCOR* could modulate the BCL-6 immune response of various cell types. In the context of *BCOR* mutations and LM it is interesting to note that affected individuals have no immunoregulatory deficits, as the *BCOR* mutation involved in LM (p.Pro85Leu) interacts normally with BCL-6, indicating that *BCOR* mutations can affect development through non-BCL-6 pathways [158].

BCOR mutations have also been implicated in a range of cancers including acute promyelocytic leukemia (APL), acute myeloid leukemia (AML), chronic myelomonocytic leukemia (CMML). Mutations include both chimeric fusion transcripts and somatic variants [154]. In patients with myelodysplastic disorders, somatic *BCOR* mutations arise secondary to mutation in genes involved in splicing or epigenetic regulation [165]. Somatic mutations are scattered through the *BCOR* coding sequence, however there is a tendency to accumulate at the c-terminal end which contains the the critical binding domains for other subunits of the BCOR complex. The p.Gly1303Ser mutation in *BCOR* lies linearly distant to the c terminal PCGF1 binding site, which begins at residue 1634. However, as previously noted, there is no structural model for BCOR at present that might provide more information on the spatial separation and effect of the residue at 1303 on binding activity at the C terminal. The proband in this case study has had ongoing unexplained dyserythropoietic macrocytosis and lowered haemoglobin levels. Considering the implications of cancer development with somatic *BCOR* mutations it would be indicated that the proband was seen by a haematologist for monitoring.

One avenue of investigation that arose from investigations into *BCOR* and a possible mechanism for the ageing disorder seen in this proband was the DNA repair defect syndrome Fanconi's anemia (FA). Although high sensitivity to cross-linking agents (chromosome fragility) is the hallmark of FA, somatic mosaicism (present in 15-25% of patients) can influence the detection of this leading to false negative test results. If reversion of a haematopoietic precursor cell occurs, then the increased proliferative advantage means that these reverted cells can expand and chromosome fragility will be reduced. Retrospective analysis of the investigations undertaken on the proband (II.3) in this study highlighted that chromosome fragility of cultured peripheral blood lymphocytes in response to exposure to mitomycin C was examined at an early age, though the result of that test did not support a diagnosis of Fanconi's anemia. The finding of skewed X-inactivation in blood (but not buccal cells or fibroblasts) of the proband raised the possibility that the X chromosome harbouring the mutant *BCOR* allele was being preferentially inactivated and therefore any investigations of chromosome fragility would be dependant on the level of inactivation of the mutant allele. Non-random X-inactivation more commonly occurs in female carriers of an X-chromosome mutation, where the mutant allele is being selected against in rapidly dividing tissues or in tissues where function of the affected gene is critical. This scenario would appear to mimic the case of reversion of a haematopoietic precursor cell leading to a normal chromosome fragility result in a patient who does in fact have an FA-causing mutation. In extreme skewing (>90:10) less than 10% of the cells in a peripheral blood lymphocyte culture will be expressing the mutant *BCOR* allele, and this ratio may be further influenced by the selective pressure on cultured cells. In this instance, the proband showed moderate skewing (88:12) that would result in <15% of cultured cells expressing the *BCOR* mutant (based on the assumption that the mutant *BCOR* allele is preferentially inactivated). The X-inactivation analysis also identified moderate skewing in the proband's mother. This was at a lower ratio (81:19), closer to the normal range seen in unaffected females (50:50-80:20), and is unlikely to be significant given that she is >50years. There is evidence for a significant increase in normal females displaying skewing in a >60 years age group compared to 28-32 years [194].

Repeated mitomycin C testing of cultured skin fibroblasts was performed on the proband after it was ascertained that these cells showed a random X-inactivation pattern. This testing also returned a negative result for FA. however, it should be noted that in a previous case of two sisters with OFCD, a positive mitomycin C test was returned for one sibling but not the other, leading to speculation that there may be confounding issues around the mitomycin C assay in the presence of *BCOR* mutations [175]. There is currently at least one example of a *BCOR* mutation identified in exome data of an FA patient [195] which may support this theory and warrant further investigation of an association between *BCOR* mutations and FA patients.

As previously mentioned in Section 4.4.5, *BCOR* mutations that are causative of OFCD are truncating mutations that are expected to produce transcripts which are degraded through nonsense-mediated mRNA decay. Hence, the mechanism of pathogenicity in OFCD is considered haploinsufficiency rather than a dominant negative effect. However the sensitivity of cells to BCOR haploinsufficiency must be considerable given that patients experience skewed X-inactivation in multiple cell types, and therefore mutant *BCOR* may only be active in <10% of cells in the first place. This does not correspond to a 50% loss of wildtype *BCOR*, but rather a 10% loss (or less). In some tissue types of course, the X-inactivation ratio is not skewed and 50% haploinsufficiency will be experienced, but these are presumably tissues in which *BCOR* function is less critical. The missense mutation seen in this case study does not truncate the transcript, and therefore should not be targeted by nonsense-mediated mRNA decay. It is reasonable to assume that mutant protein would be produced and expressed in cells, and could possibly be altering the normal function of *BCOR*, in contrast to the haploinsufficiency effect seen in OFCD. As the proband in this case study shows moderate-extreme skewing in at least one tissue type analysed, the ratio of mutant *BCOR* that could potentially be expressed ranges from <10% to 50%. This could lead to differing effects depending on the skewing present a given tissue or cell type. It would be interesting to quantify expression of mutant *BCOR* in differing tissues, particularly those with apparent random X-inactivation to determine if expression is truly 50% or there is some bias against correct processing of the mutant RNA. Unfortunately, acquiring such tissues would be highly invasive and was not considered necessary for the project.

Difficulties in pursuing this *BCOR* variant further in the setting of this affected individual include lack of structural information, and a clear understanding of how components of the BCOR complex associate. Without understanding the importance of the affected sequence region (and how this may impair complex association by structurally preventing association or altering a recognition sequence) it is difficult to explain how a mutation in *BCOR* could lead to the phenotype observed. Further work is needed to properly explore the context of this variant in this individual, however, such work is outside the scope of this PhD thesis.

Chapter 5

A case of sudden cardiac death

5.1 Family Christchurch

This chapter focuses on an undiagnosed family with an inherited cardiac phenotype. This chapter discusses the analysis of the exome sequencing data, while subsequent chapters will explore the research avenues that have arisen from the findings in this chapter.

5.1.1 Case history

A Caucasian family of two parents and four children were identified as harbouring a possible genetic abnormality following the death of two teenage male children (II.1 and II.3 in Figure 5.1) with presumed sudden arrhythmia (Figure 5.1). On autopsy, marked mid-myocardial fibrosis was documented. Subsequent clinical review of the surviving children (II.2 and II.4 in Figure 5.1) detected dilated cardiomyopathy with mid-myocardial fibrosis. All four children in this family had previously shown an unexplained, exquisite sensitivity to alcohol ingestion resulting in pain and discomfort. Analysis of the parents showed no cardiac abnormalities and they were considered unaffected. No other patients with this set of features have been identified to date and it is believed the family may characterise a novel syndrome with predisposition to sudden arrhythmic cardiac death.

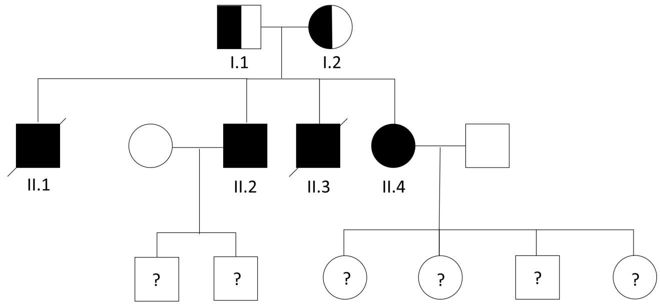


Figure 5.1 Family Christchurch pedigree showing cardiac condition
All four children are affected with the familial cardiac condition as diagnosed by cardiac MRI (CMR). Individuals II.1 and II.3 died suddenly at ages 15 and 20 years respectively. Both parents are unaffected on cardiac MRI (CMR).

Individual II.1 collapsed and died aged 15, after drinking a small amount of beer (150mL, or 6g of ethanol). He was previously well, and had no prior cardiac symptoms, but like all his siblings had been exquisitely sensitive to alcohol. This manifested as dull pain in the chest, a feeling of heaviness throughout the limbs, and pallor, following the consumption of small amounts of alcohol (typically dessert such as trifle or medicine such as children's cough syrup). At post mortem examination the only macroscopic abnormality observed was slight dilation of the heart, with both ventricles being enlarged and small pale area on the epicardium of the left ventricle. Microscopic examination of cardiac tissue revealed evidence of focal inflammation with neutrophils, lymphocytes and eosinophils. The coronary arteries were normal. A diagnosis of myocarditis and sudden arrhythmic cardiac death was made.

Individual II.2 exhibited the same sensitivity to alcohol. He was clinically assessed because of his family history after the death of sibling II.1. Physical examination was normal as was an echocardiogram, exercise test and Holter monitor. However a MRI scan showed marked mid-myocardial fibrosis (Figure 5.2). He subsequently received an implanted defibrillator for primary prophylaxis of sudden arrhythmic cardiac death. No events of device triggering have occurred to date.

Patient II.3 died aged 20 years after drinking a small amount of alcohol (anecdotally, two spirit drinks). He was previously well, and had no prior cardiac symptoms, but had also been

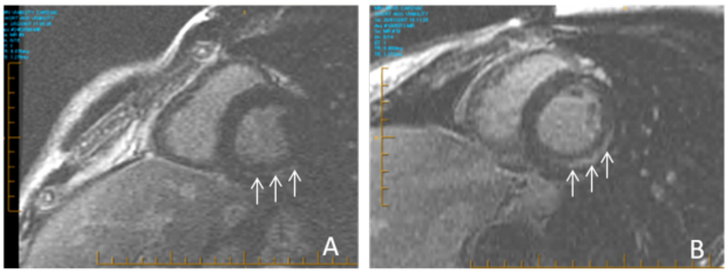


Figure 5.2 Cardiac MRI
Siblings II.2 and II.4 show prominent midmyocardial fibrosis on CMR.
A Sibling II.2 shows moderate cardiac fibrosis.
B Sibling II.4 shows severe cardiac fibrosis.
Fibrosis marked by arrows. Image credit: Ian Crozier.

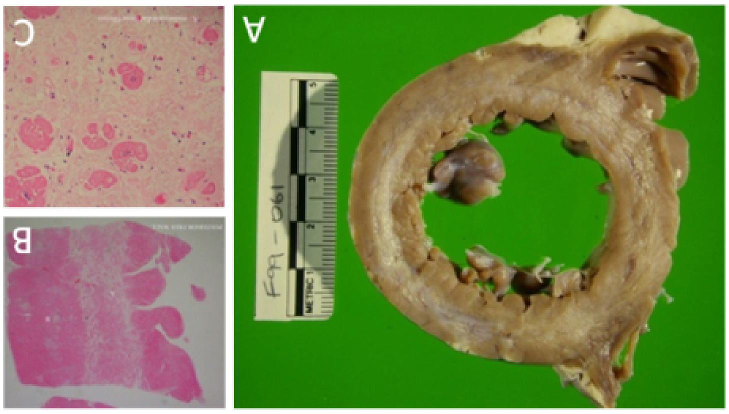


Figure 5.3 Heart section
A Sibling II.3, post-mortem transverse section through left ventricle showing a virtually circumferential lamina of scarring in midmyocardium with focal subendocardial involvement.
B / **C** Low and high power microscopy of the posterior free wall of the left ventricle showing prominent midmyocardial loose fibrosis. Image credit: Ian Crozier.

exquisitely sensitive to alcohol. At post mortem examination the only abnormalities observed were in the heart. The heart weighed 395g (normal 300g). The left ventricle was dilated with a virtually circumferential lamina of scarring in the mid-myocardium and with focal subendocardial involvement. Microscopic examination revealed very widespread, mostly mature, scarring of mid-myocardium in all sectors (Figure 5.3). No ischaemic changes were observed and there was no microscopic evidence to suggest acute hypersensitivity or interstitial acute myocarditis. The coronary arteries were normal.

Patient II.4 remains well with no cardiac symptoms, but exhibited the family sensitivity to alcohol. She was clinically assessed because of her family history after the death of her sibling; individual II.1. Physical examination was normal, as was echocardiogram, exercise test and Holter monitor. However a MRI showed marked mid-myocardial fibrosis (Figure 5.2). She subsequently also received an implanted defibrillator for primary prophylaxis of sudden arrhythmic cardiac death. No events of device triggering have occurred to date. At the time the defibrillator was implanted a skeletal muscle biopsy was obtained which showed no obvious abnormality under histopathology examination, although a muscular dystrophy or metabolic disorder could not be excluded. Stains for fat and glycogen were within normal limits. A panel of enzyme antibody stains (including myophosphorylase, NADH, COX, SDH, MADA, aldolase and PFK) was normal. Immuno-histochemical studies of skeletal muscle showed changes suggestive of a mild chronic myopathy. However immunostaining for dystrophin, dysferlin, emerin and laminin showed no obvious abnormality. Urine organic/amino acid, and blood spot acylcarnitine profiling showed no abnormalities. Whole blood carnitine was within normal limits at 23µmol/L (normal range 11-58).

5.1.2 Array CGH

In order to detect significant copy number variants within the family, which wouldn't be detected by DNA sequencing approaches, arrayCGH was performed using Nimblegen 135k oligo microrrays in the four living individuals. No clinically significant CNVs were detected in any of the four family members (data not shown), nor were there any variants of unknown clinical significance. We detected common CNVs which segregated within the family. Although not exhaustive, this analysis discounted larger CNV's (>50kb) underlying the inheritance of this pedigree's cardiac condition.

5.1.3 Targeted genetic analysis

Previous molecular investigation in II.4 included targeted sequencing of all 12 exons of the lamin A/C gene (LMNA) and also all coding regions of the lamin-associated protein 2 gene (LAP2). These exons, together with flanking intronic sequences were analysed and no mutations were detected. Mitochondrial genome sequencing was undertaken on whole blood, buccal cells and a skeletal muscle biopsy from II.4 and on whole blood from II.3. No known pathological mitochondrial sequence variants were detected. However both siblings were found to be homoplasmic for a novel sequence variant (m.9751T>C) in the Cytochrome C oxidase subunit III (MT-CO3) gene, which results in the substitution of a phenylalanine for a serine at residue 182 of the COXIII protein (p.COXIII:Phe182Ser) (Data not shown). Other mutations in MT-

CO3 are known to be causative of mitochondrial disorders such as Leigh syndrome, Leber's optic atrophy and cytochrom-c oxidase deficiency. However sequence analysis of the MT-CO3 gene from a whole blood sample from the mother showed she was also homoplasmic for the m.9751T>C sequence variant, indicating the substitution is unlikely to be solely responsible for the phenotype seen in her children.

Due to the lack of adequate explanation for the alcohol sensitivity and sudden cardiac death in this family, further analysis was undertaken. Instead of sequentially sequencing less convincing candidate genes in an attempt to uncover a basis for the condition, the implementation of exome sequencing allowed us to investigate the entire coding region of the genome, avoiding the bias of selecting genes to be analysed.

5.1.4 DNA preparation

DNA from all six family members had been previously extracted for mitochondrial sequence studies, and stored at 4°C for >5 years. DNA from I.1, I.2, II.2 and II.4 was in usable condition, however, samples from II.1 and II.3 were considered fully degraded and not suitable for further analysis (data not shown).

5.1.5 Exome analysis

Library preparation and exome sequencing was performed by New Zealand Genomics Limited (NZGL). The Illumina software enrichment pipeline v.0.9 generated between 64,691 - 71,555 annotated variants for each individual exome, this data was returned in variant call files (VCF). Further filtering of the VCF's in ANNOVAR was performed to exclude variants with a MAF>0.01. This reduced the number of variants to 3,350 - 6,581 variants which were available in VARLIST files. After manual filtering within excel to exclude variants not shared by the two affected siblings, the number of variants on which to focus further attention on was reduced to 724.

Cardiomyopathy-associated gene analysis

A comparison of these 724 variants to a list of 41 cardiomyopathy-associated genes (See Table 5.1) yielded three variants in the large gene TTN that were common to both affected siblings. These three variants were paternally inherited, and polymorphic. No other genes from the candidate list had variants present in both affected siblings. Titin is the largest known protein with up to 364 exons , and missense variants of no pathogenic consequence are very common, with an average of 23 present in any individual in the EVS database [196]. No variants from this panel were progressed for further analysis, and a wider analysis of the exome data was considered.

Table 5.1 Cardiomyopathy-associated gene list

ABCC9	ILK	PDLIM3	TPM1
ACTC1	LAMP2	PLN	TTN
ANKRD1	LAP2	RBM20	TTR
ATCN2	LDB3/ZASP	SCN5A	TXNRD2
BAG3	LMNA	SGCD	VCL
CRYAB	MYBPC3	TAZ	PSEN2
CSRP3	MYH6	TCAP	PSEN1
DES	MYH7	TMPO	LAMA4
DMD	MYPN	TNNC1	
EMD	NEBL	TNNI3	
EYA4	NEXN	TNNT2	

Exome wide analysis

As both parents of these affected siblings are unaffected, a recessive model of inheritance was explored. Fifty genes were identified as containing two or more variants in both living affected siblings. Of these, 47 were excluded from further exploration due to lack of evidence supporting relevance to this family's phenotype (Figure 5.4). Three genes remained as potential candidates for a cardiac-type disorder; *TTN* (Titin), *KCNJ12* (Potassium Voltage-Gated Channel Subfamily J Member 12) and *PPA2* (Inorganic Pyrophosphatase 2). *KCNJ12* encodes a protein which functions as an inwardly rectifying channel contributing to the cardiac inward rectifier current. *TTN* has previously been associated with familial hypertrophic cardiomyopathy, and

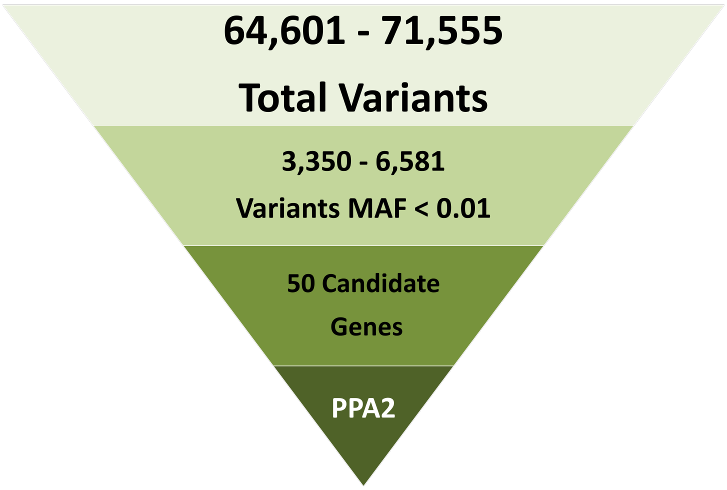


Figure 5.4 Variant filtration
Schematic of variant filtration, showing the scale of variants detected by whole exome sequencing (WES).

as such was included on the candidate gene list in the approach above (see Section 5.1.5). *PPA2* produces an enzyme involved in the mitochondrial oxidation-phosphorylation pathway (OXPHOS) which has multiple components implicated in cardiac myopathy [197]. Analysis of all these genes was performed in both unaffected parents. The three TTN variants were previously excluded by the candidate gene approach as being paternally inherited and polymorphic while the two KCNJ12 variants were excluded based on non-segregation with the phenotype.

The remaining candidate gene, *PPA2*, had not been associated with human disease previously; however the localisation of the product to the mitochondrial OXPHOS pathway was of interest. The two variants identified in *PPA2*; c.514G>A (p.Glu172Lys) and c.683C>T (p.Pro228Leu) segregated with the phenotype, and importantly, both parents were found to each have one of the two variants suggesting a compound heterozygous mode of inheritance if proved pathogenic (Figure 5.5). Therefore, *PPA2* was the sole remaining candidate following filtering. The fact that *PPA2* is a mitochondrial-associated gene meant that the additional *COXIII* variant identified in all four children (see 5.1.3), was not excluded as a factor that could be contributing to the familial condition.

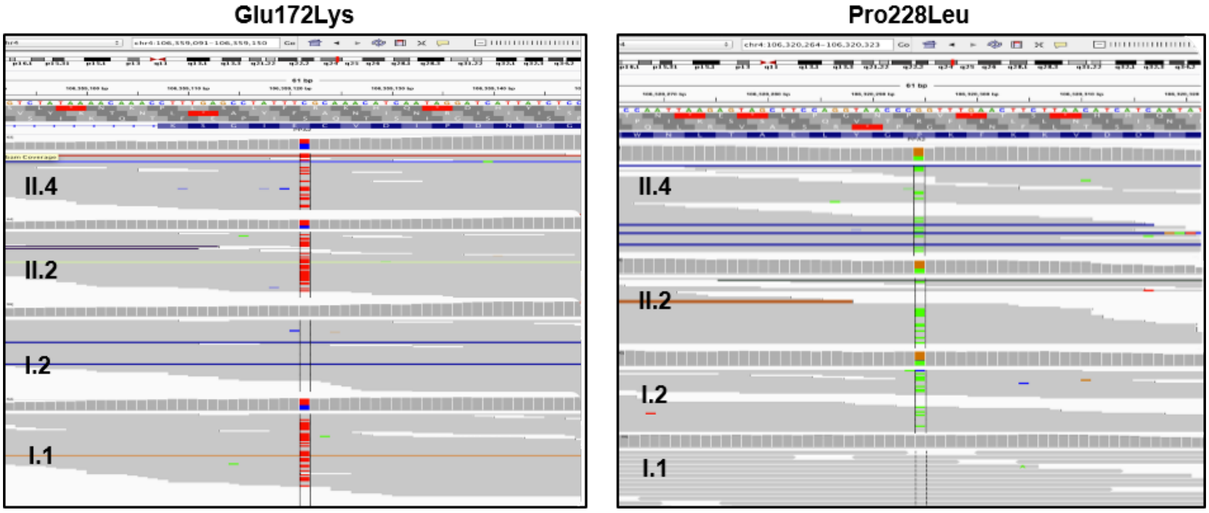


Figure 5.5 PPA2 variants - IGV visualisation
Variants c.514G>A (p.Glu172Lys) (left) and c.683C>T (p.Pro228Leu) (right) can be seen in both affected surviving children (II.4 and II.2). Variant p.Glu172Lys is present in the the father (I.1) but not mother (I.2). Conversely, p.Pro228Leu is present in the mother (I.2) but not father (I.1) confirming children are compound heterozygous for these variants.

5.1.6 Sanger sequencing

Although initial analysis was performed on historical DNA (including FFPE autopsy samples), new EDTA blood samples were collected from the living patients for DNA extraction and confirmatory Sanger sequencing. DNA from stored Guthrie spots was used in the case of patients II.1 and II.3. Sanger sequencing of the region surrounding p.Glu172 and p.Pro228 confirmed that all four children were compound heterozygotes for both mutations, located in exons 6 and 8 of *PPA2* (Figure 5.6). Supporting this, both parents were confirmed to be carriers of one mutation (Glu172Lys = paternal; Pro228Leu = maternal).

5.2 *PPA2* variant investigations

5.2.1 *In silico* predictions of *PPA2* variants

The *PPA2* variants; c.514G>A (p.Glu172Lys) and c.683C>T (p.Pro228Leu) were assessed *in silico* for predicted functional effects and evolutionary conservation of the affected residue. These predictions strongly indicated that both variants involve highly evolutionarily conserved residues, and are predicted to cause severe functional effects (Tables 5.2). Glu172Lys is actually present in the database dbSNP (rs146013446), however, given the segregation data and damaging *in silico* predictions it was not excluded based on this.

Table 5.2 *In silico* predictions for *PPA2* variants

	SIFT	PolyPhen-2	MutationTaster	Grantham	MAF
p.Glu172Lys	0	1	1	56	$4.9E^{-4}$
p.Pro228Leu	0	0.996	1	98	$2.5E^{-4}$

Damaging (Pink); Moderate (Orange); Neutral (Green)

5.2.2 Modelling

The high degree of homology between the human and yeast pyrophosphatase proteins allowed predictive modelling based on the known yeast structure of PPA1 [198] in the absence of a PPA2 model. This allowed visualisation of the predicted change in hydrogen bonding and structural characteristics of the variant residues compared to wildtype at the two points of interest. A substitution of glutamine to lysine at residue 172 is predicted to disrupt at least 3 hydrogen bonds between interacting protein chains (Figure 5.7). A substitution of Proline to Leucine at residue 228 is also predicted to disrupt the secondary structure of PPA2 based on the exceptional conformational rigidity that proline confers. This residue forms the turn between two helixes, and may be essential for the correct position and orientation of these structures. Proline's struction mean that in evolutionary terms it forms an important element of the tertiary protein structure, allowing twists and turns of the amino acid sequence to facilitate distinct domain formation and transitions from α to β sheets (Figure 5.8).

5.3 *In silico* prediction of mitochondrial variant

Previous *in silico* analysis of the COXIII variant m.9751T>C (p.Phe182Ser), indicated that the phenylalanine residue is not highly conserved among vertebrates. There is, however, some structural conservation (Figure 5.9). Phenylalanine, tyrosine and less commonly isoleucine are

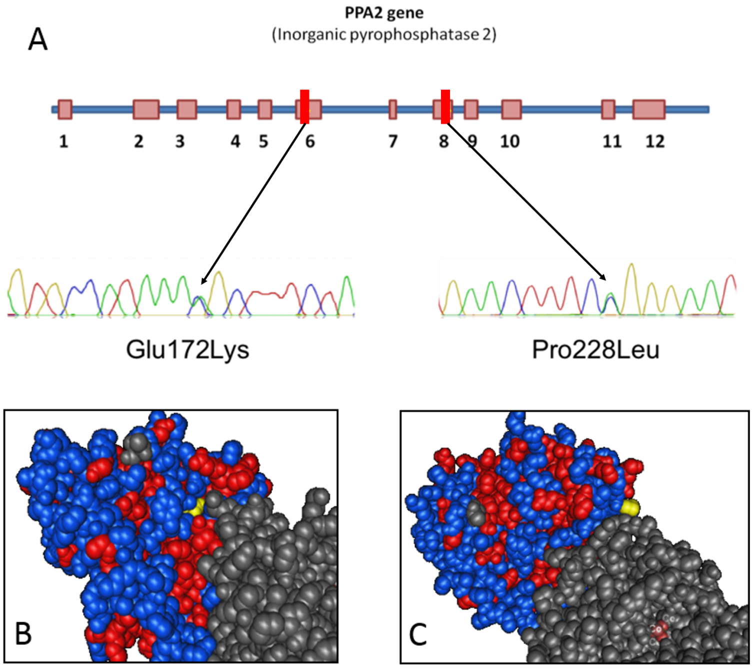


Figure 5.6 Variant resequencing and modelling.
A Variant re-sequencing and exon position of *PPA2* variants p.Glu172Lys and p.Pro288Leu. Electropherogram shown is an example of one affected family member who is compound heterozygous for the two PPA2 variants. **B** *Saccharomyces cerevisiae* space fill model indicating the position of the homologous Glu172 residue in yellow. **C** *Saccharomyces cerevisiae* space fill model indicating the position of the homologous Pro228 residue in yellow.

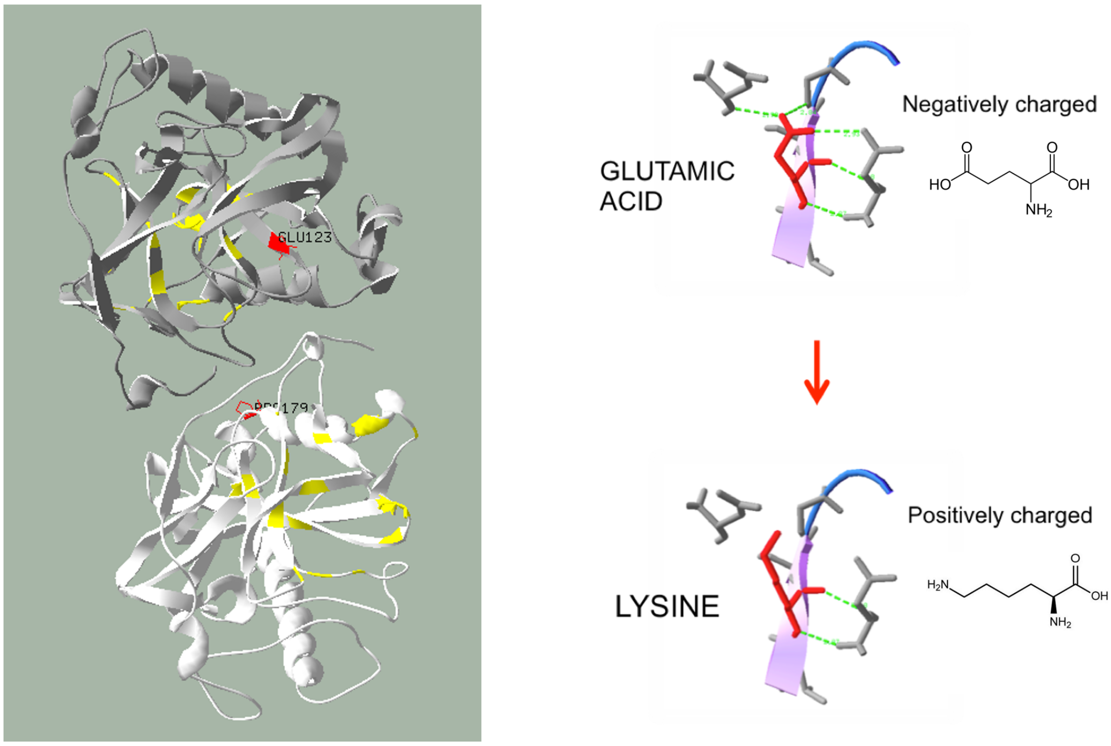


Figure 5.7 Yeast structural model (PPA1) showing predicted position of human PPA2 variant at 172
Position within dimer molecule of residue 172 highlighted in red (Glu123 label refers to equivalent residue in Saccharomyces structure). Putative active site residues are highlighted in yellow (Left). A substitution of Glutamine to Lysine at residue 172 is predicted to disrupt at least 3 hydrogen bonds between interacting protein chains, and convert the charge on the residue from negative to positive (Right)

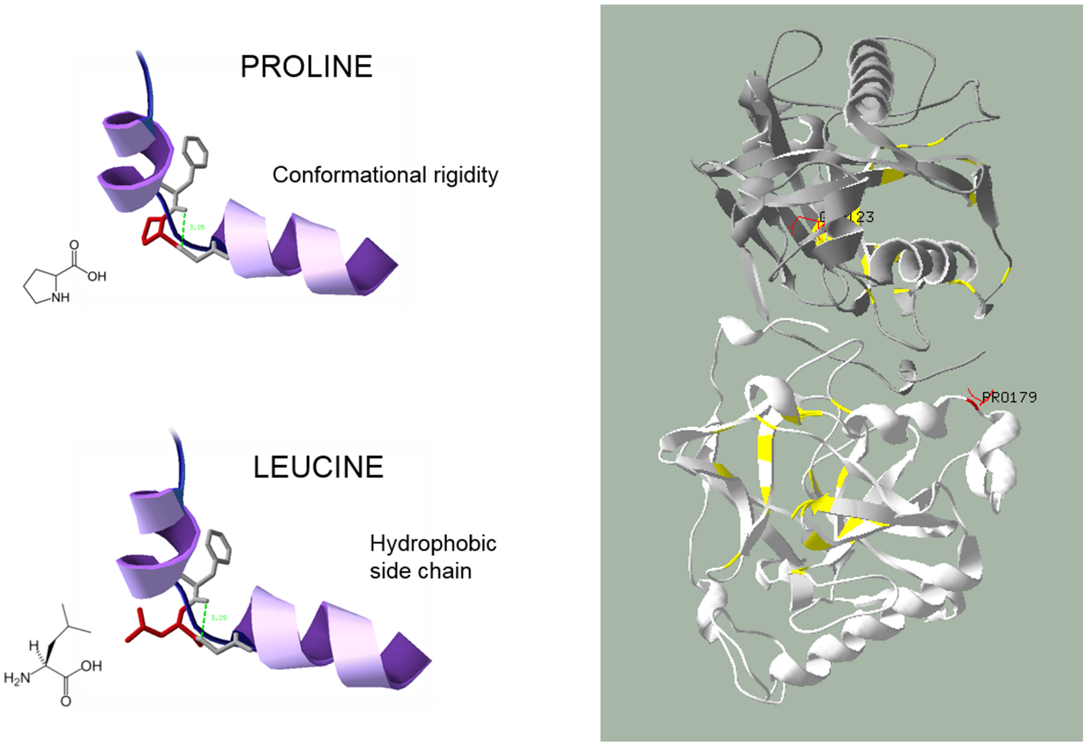


Figure 5.8 Yeast structural model (PPA1) showing predicted position of human PPA2 variant at 228.
A substitution of Proline to Leucine at residue 228 is predicted to disrupt the secondary structure of PPA2 based on the exceptional conformational rigidity that proline confers (Left). Position within dimer molecule of residue 172 highlighted in red (Pro179 label refers to equivalent residue in Saccharomyces structure). Putative active site residues are highlighted in yellow (Right).

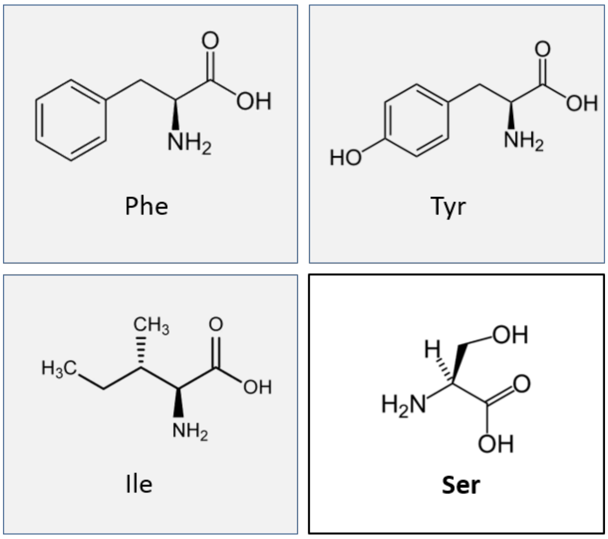


Figure 5.9 COXIII structural conservation
Structural conservation between the amino acids present at COXIII residue 182. Phenylalanine is the wildtype human residue, whilst tyrosine and infrequently isoleucine are present in different species. These three amino acids all contain a hydrophobic side chain, and are physiochemically quite distinct from the Serine residue seen at position 182 in the maternal line of this family.

the only three amino acids present across an extensive range of sequence entries. These amino acids all contain a hydrophobic side chain and are grouped similarly under several measures of amino acid diversity. Given the structural conservation observed this residue may be more damaging than previously suggested by *in silico* algorithms.

5.4 Discussion

Mitochondrial sequencing and whole exome sequencing (WES) have been used to extensively investigate the genetic basis of this familial condition, identifying three variants that are considered strong candidates for pathogenicity. Two of the variants in the nuclear encoded (but mitochondrial associated) gene *PPA2* (p.Glu172Lys and p.Pro228Leu) were identified segregating in a compound heterozygous configuration in two symptomatic individuals in a family with a history of dilated cardiomyopathy and sudden cardiac death. Sanger sequencing confirmed the presence of both variants in all four affected children, and that both parents were carriers of one variant. *In silico* analyses supported the hypothesis that these variants would be damaging, despite p.Glu172Lys being included in the dbSNP database as a polymorphism (rs146013446).

As previously detailed, in addition to the *PPA2* variants, the affected individuals (and their mother) are homoplasmic for a novel sequence variant (m.9751T>C) in the Cytochrome C oxidase subunit III (MT-CO3) gene, which results in the substitution of a phenylalanine to a serine at residue 182 of the COXIII protein (p.COXIII:Phe182Ser). This variant was classified as of unknown significance when it was first identified, and early *in silico* prections indicated it may be benign. When analysing this variant in the context of *PPA2* variants also being present, the low PolyPhen-2 score was attributed to an apparent lack of conservation of residue Phe182 within MT-CO3. In a comparison of this residue across multiple species, phenylalanine, tyrosine and isoleucine are all represented but not serine, notably. So, although there is weak sequence conservation, there is, strong structural conservation. Phenylalanine, tyrosine and isoleucine all share chemical features that are distinct from the serine residue which is present in this family. Bearing this in mind this variant may have more functional importance than the *in silico* measures initially predicted, with the distinct possibility that it may genetically interact with the *PPA2* variants.

Modelling of the *PPA2* variants based on the highly homologous yeast pyrophosphatase structure indicated the potential importance of their position within the molecule. Glu172 lies directly behind the magnesium binding catalytic cleft. Previous studies in yeast have shown that altering residues around Glu172 can completely abolish PPA2 function [199]. In particular residues 164 and 169 are necessary for the binding of magnesium [199]. In contrast Pro228 is not located within the active site of the molecule, but lies on the interface of the two PPA2 molecules interacting to form a dimer. As a proline residue confers exceptional conformational rigidity and peptide chain turning ability, it would be predicted that a Pro228Leu mutation could disrupt this interface structure. It is interesting to note that given the dimeric nature of PPA2 molecules, and the presence of two independent mutations, there are three different dimer products that can potentially form a compound heterozygote (Figure 5.10). The difference between a carrier and a compound heterozygote is the overall complement of wildtype PPA2 dimers, dimers with one mutant molecule, and dimers with two mutant molecules (either the same mutation or different). In a carrier, 1/3 of dimers would be wildtype (two wildtype

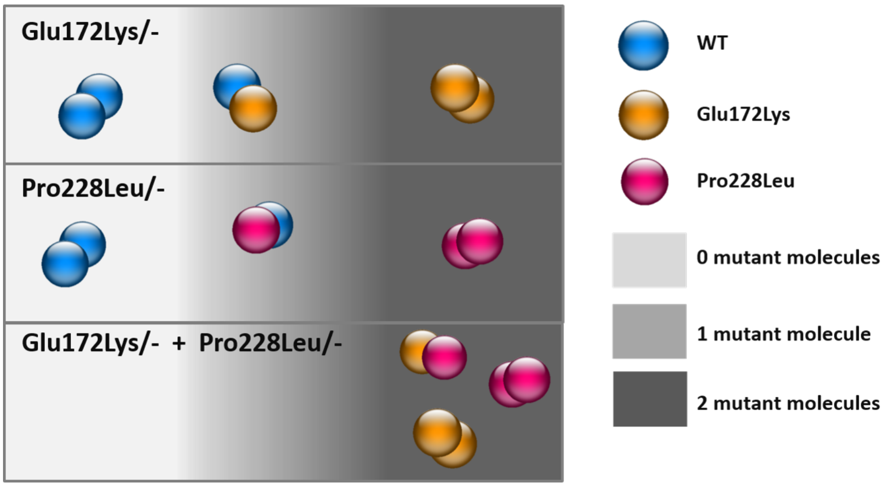


Figure 5.10 Dimer combinations

Combinations of dimer molecules that can occur in *PPA2* mutation carriers, and compound heterozygotes.

molecules). In a compound heterozygote 0/3 of dimers would be wildtype. It would be interesting to compare the functionality of these dimer products and determine approximate percentage of PPA2 function in a carrier vs a compound heterozygote setting.

The Exome Aggregation Consortium (Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org [Sept 2016]) browser is currently one of the most comprehensive summaries of exome data; consisting of over 120000 high-quality allele calls from 60706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Cases of severe paediatirc disease have been removed, so this data set should serve as a useful reference set of allele frequencies for severe disease studies.

The allele frequency of the p.Pro228Leu variants in ExAC is <0.0002 (30/120268 alleles).The allele frequency of the p.Glu172Lys variants in ExAC is <0.0004 (59/120800 alleles) Interestingly, of all 147 missense and loss of function (LOF) *PPA2* variants included in the ExAC database, only four missense mutations have been observed in a homozygous state, and one of those has been observed only once. This is usually a reflection of strict conservation required

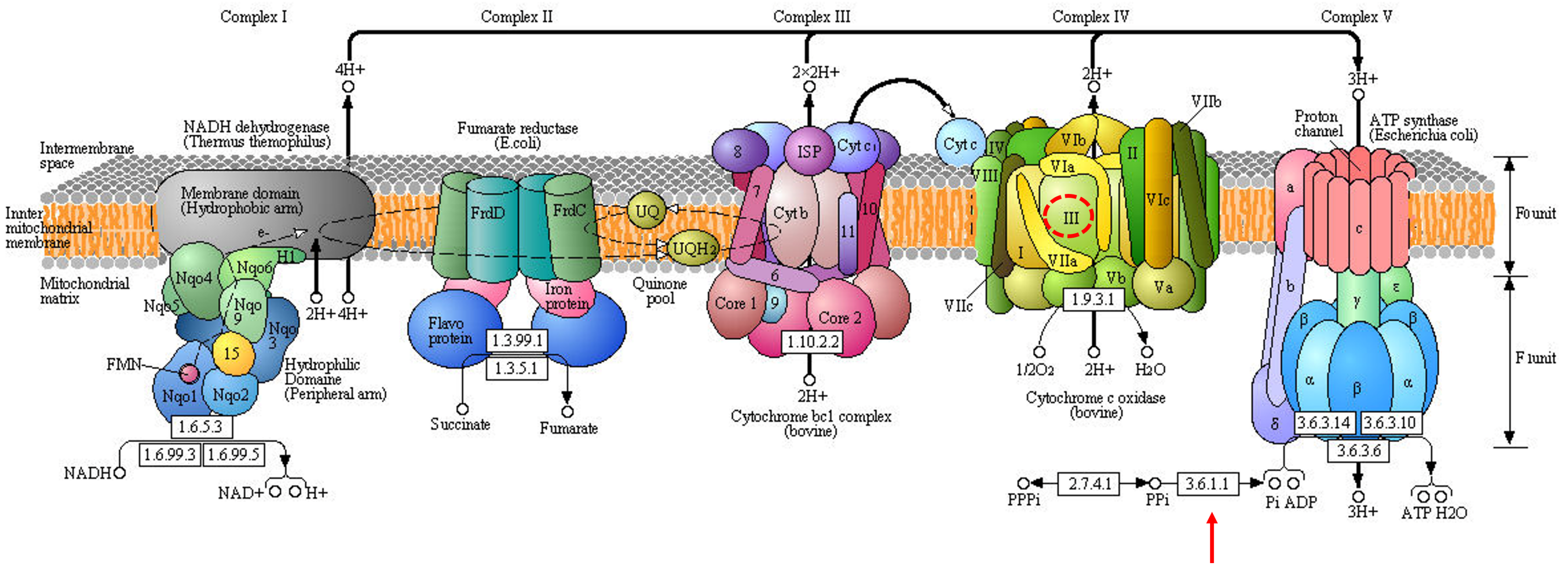


Figure 5.11 OXPHOS schematic
Scheme of Mitochondrial oxidative-phosphorylation (OXPHOS) complexes for the generation of ATP. The complex IV subunit encoded by MT-CO3 is indicated by the white box. Inorganic pyrophosphatase (PPA2) catalyses the process 3.6.1.1 indicated by the arrow. Reproduced with permission from the Kanehisa Laboratories and the KEGG project (www.kegg.org).

for correct function of the protein. Variants that either do not occur at all or are not present in the population in a homozygous state reflect strict functional control at the coding region of the genome. No occurrence of homozygosity was observed for either p.Glu172Lys or p.Pro228Leu.

The occurrence of the Glu172 Lys variant in dbSNP is important, as many filtering strategies exclude all variants present in this database as a way of minimising polymorphisms in the variants to be investigated. Applying such a filtering strategy in this case would not have brought *PPA2* to attention. As quoted by Sherry *et al* "There is no requirement or assumption about minimum allele frequencies or functional neutrality for the polymorphisms in the database. Thus, the scope of dbSNP includes disease-causing clinical mutations as well as neutral polymorphisms" [200]. Therefore the use of presence of dbSNP as exclusion criteria should be avoided. The possible involvement of alcohol in relation to the SCD events occurring in this family is a unique factor that warrants further investigation. There are known reactions to alcohol metabolism defects, such as the *ADH* and *ALDH* mutations which contribute to the "flushing reaction" seen frequently in Asian populations [201], though no variants of this type were observed in affected members of this family. The effects apparently induced by alcohol ingestion in these individuals were described as a generalised but severe aching in their torso and limbs, accompanied by pallor. Both SCD events occurred in the hours after alcohol ingestion possibly suggesting alcohol is triggering SCD in this family, although it is not clear how this relates to the mutations in *MT-CO3* and *PPA2* observed.

Mitochondria perform a range of tasks such as the Krebs cycle, pyruvate oxidation, amino acid, fatty acid and steroid metabolism, and generation of ATP via the respiratory chain. The respiratory chain is also know as the electron transport chain/oxidative phosphorylation system (OXPHOS) (See Figure 5.11). The system is comprised of five protein complexes in the inner mitochondrial membrane, and involves two small electron carriers; ubiquinone (coenzyme Q10) and cytochrome c (Figure 5.11). There is extensive evidence that mutations that impair OXPHOS function, reducing energy supply to the heart and consequently resulting in cardiomyopathy [202]. Complexes I to IV are enzyme complexes, while complex V is the ATP synthesis unit. Energy generated by electron transport across the four enzyme complexes drives the ATP production in complex V [203, 204, 205].

The subunits of all these complexes are encoded in both nuclear DNA and mtDNA. Defects within each complex have their own array of clinical effects which are well described [206, 207]. Complex IV comprises 13 subunits, 10 of which are nDNA encoded. The catalytic core of the complex is comprised of the three mtDNA encoded subunits (COXIII). There are no known mutations that cause disease in the 10 nuclear encoded genes, however defects in COXIII can cause phenotypes such as Leigh Syndrome, fatal infantile COX deficiency, hypertrophic cardiomyopathy and myopathy[203]. Complex V comprises 16 subunits of which 14 are nuclear encoded and do not have known pathogenic mutations to date. Complex V defects (in the mtDNA encoded subunit) are, however, associated with Leighs syndrome and a syndrome of neuropathy, ataxia and retinitis pigmentosa (NARP) [208]. Although respiratory chain enzyme activities were found to be normal in skeletal muscle from sibling II.4, complex V activity was not measured as the reaction conditions (which include phosphate buffers) were not explicitly designed to test this activity and would artifically provide free phosphate. It is interesting to note that that the COXIII variant observed in this family is present on a background of the mitochodrial haplogroup U, previously shown to increase the severity of cardiomyopathies [209, 210].

Mitochondrial disorders have been extensively discussed in the literature, but there are no previous reports on *PPA2* associated with human disease. A study in yeast [199] has shown that the pyrophosphatase encoded by *PPA2* is essential for mitochondrial DNA maintenance. When PPA2 function was disrupted the strain was viable, but could no longer grow using respiratory carbon sources such as glycerol, acetate, or ethanol. Fluorescent microscopy showed that these cells appeared to be a p^0 strain, completely lacking mitochondrial DNA. However, a study of *PPA2* in a cohort of patients with mtDNA depletion syndromes failed to show any pathogenic mutations, and provided no evidence for *PPA2* involvement in mitochondrial depletion in humans [197].

All three variants identified in this family affect proteins that are located in the OXPHOS pathway within the mitochondria, specifically one mtDNA variant within complex IV (COXIII), and two compounding variants within *PPA2*, which functions to hydrolyse pyrophosphatase (Pi-Pi) to individual phosphate molecules (Pi) for use in various catalytic processes including ATP generation at complex V, the ATP synthesis unit. Further clarification of the role of PPA2 within the mitochondria is required to understand the possible disease mechanism(s) involved in this pedigree. Identifying other families with *PPA2*mutations and OXPHOS dysfunction was considered a priority to fully establish a causative role of *PPA2* in this cardiac disorder.

s

Chapter 6

Wider investigations of *PPA2*

6.1 Intro

The previous chapter discussed the discovery of two variants in the gene *PPA2* which were considered putative causal mutations resulting in a cardiac phenotype seen in our index family. Following this discovery, we approached collaborators who had access to a database of exome data from patients with mitochondrial disease. After re-examining their data, and approaching other laboratories with similar cases, a further three families with *PPA2* mutations were identified based on genotype (not clinical features). These families represent an additional six patients, bringing the total of affected patients including our index family to ten. This chapter will discuss these additional cases, and our efforts to demonstrate pathogenicity. We also aimed to link the varying phenotypes seen between the four affected families to *PPA2* mutations of varying pathogenicity, and more theoretically propose a potential disease mechanism for *PPA2* dysfunction.

6.2 Case details

The affected individuals in each family are shown in Figure 6.1, including the index cases which will be herein referred to as Family Christchurch. Individuals P5-P7 (Family Basel) were born to consanguinous parents of Sri Lankan origin and affected children were homozygous for the

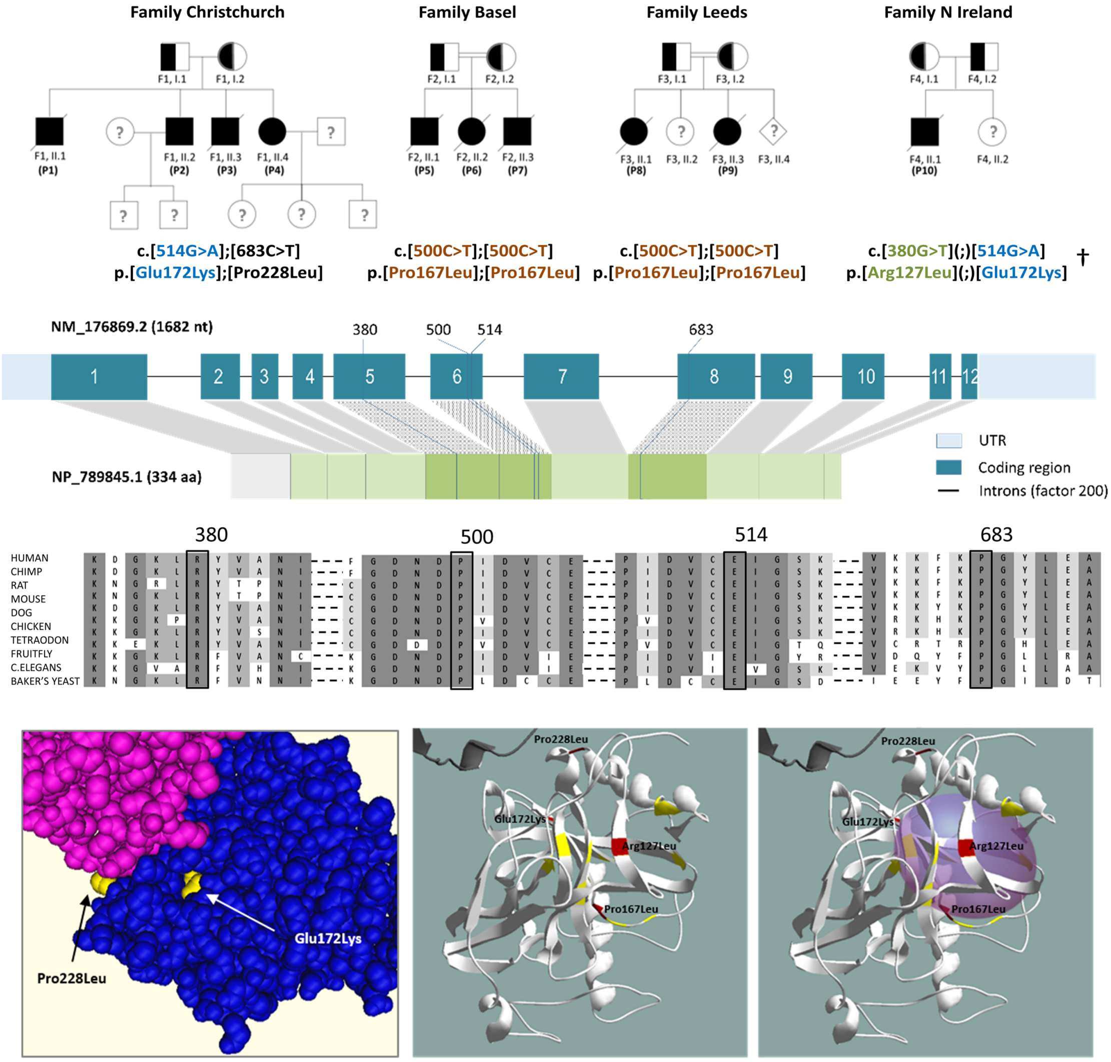


Figure 6.1 PPA2 pedigrees
A Pedigrees of four families identified with mutations in *PPA2* (GenBank NM 176869.2) encoding the mitochondrial inorganic pyrophosphatase.**B** Location of mutations within the gene, and phylogenetic conservation of the predicted missense mutations. **C** Space fill model showing position of p.Pro228 at boundary of dimers and p.Glu172 in the active site. **D** Left: Structural model of one molecule of PPA2 showing the position of 4 mutations in folded structure (red). Residues that are known to be critical to PPA2 function in S. cerevisiae are highlighted in yellow. Right: Space fill of the PPA2 active site showing three mutations are located at the surface of the active site.

PPA2 mutation c.500C>T (p.Pro167Leu). Individual P5, a boy, was born spontaneously after an uneventful pregnancy. Birth weight was 2,820g, length 49 cm, head circumference 33.5 cm, and Apgar scores 9/10/10. The first days of life were uneventful and mother and child left the hospital on day six. On the 11th day of life the child was readmitted to hospital with signs of tonic clonic seizures and vomiting. The patient was hypotonic, somnolent and pale, though heart and lung function seemed normal. Lactate was elevated at 10.5 mmol/l (normal 0.4-2.8 mmol/l). In the following days he again had tonic-clonic seizures, which could be suspended temporarily by treatment with Diazepam and Phenobarbital. Oxygen saturation was persistently low necessitating intubation. Respiration did not improve requiring artificial ventilation. Generalised tonic clonic seizures persisted. A subsequent ECG investigation showed a convex ST segment elevation and the child died hours later with severe bradycardia. Investigation of plasma amino acids showed elevated alanine. Investigation of organic acids was normal. Investigation of heart autopsy revealed fresh myocardium necrosis, mainly of the right heart, and interstitial lymphocyte infiltration. Electron microscopy of the heart showed mitochondria with degeneration of cristae but no evidence of viral infection.

Individual P6, a girl, was born after a normal pregnancy and spontaneous delivery with good postnatal adaptation. At the age of 14 days, two hours after an unremarkable routine check by a paediatrician, the child suddenly deteriorated and was admitted with marked tachypnea. The child had vomited once and showed generalized seizures. Muscle tone was slightly hypotonic. In the intensive care unit a marked metabolic acidosis with a blood pH of 6.9 (normal 7.35-7.45), HCO3 at 4 mmol/l (21-26 mmol/l), lactate 22 mmol/l, and pyruvate 253 mmol/l (normal 84-784) was recorded. In the following hours there was a cardio-respiratory decompensation and the girl died 6 hours after the onset of symptoms. Autopsy revealed acute and sub-acute necrosis of the myocardium in both the left and right heart, though more prominent in the right heart. Electron microscopy of the heart showed mitochondria with degeneration of cristae as observed in her brother (P5). Furthermore, multiple sub-acute necroses were found in both cerebral hemispheres. Investigation of the respiratory chain enzymes and pyruvate dehydrogenase were normal in skeletal muscle and fibroblasts.

Individual P7, a boy, was born at term after a normal pregnancy. He was hospitalized in intensive care unit from the first minute of life and carefully observed due to the family history. He was perfectly healthy in his first days of life, similar to his siblings. Assuming that the siblings might have suffered from a defect in the respiratory chain isolated to the cardiac muscle this patient was supplemented with a cocktail of vitamins usually given in defects of the respiratory chain (coenzyme Q10, riboflavin, vitamins C, E, carnitine, biotin, beta-carotene). In the following days he exhausted during feeding and developed signs of slight cardiac failure. Selective screening of inborn errors did not reveal any pathological findings. Plasma lactate remained in the normal range. On day nine, the clinical symptoms reminded a clinician of a nutritional thyamine deficit, and the patient was given intravenous thiamine hydrochloride 20 mg/d, which markedly improved his condition. Carnitine, vitamin K, biotin and coenzyme Q10 as well as thiamine were discontinued on day 11 since the patient seemed to be perfectly well. However, his condition worsened again and heart failure became evident with occasional arrhythmia. Levels of troponin and transaminases increased. Echocardiography showed impaired function of the enlarged right ventricle. On day 15 it was decided to supplement him with thiamine 3 x 100mg/day orally. Carnitine, vitamin K, biotin and Q10 were reintroduced. Heart function improved, and the troponin and transaminase level normalized. However, on day 17 recurring tachycardia occurred, which responded temporarily to adenosine and then to electroconversion, but it recurred in the following days. With 5-20 mg i.v. thiamine, a regular sinus rhythm could be obtained and the patient improved dramatically. However, in spite of thiamine 30-80 mg i.v. daily and propafenone, severe arrhythmia (Hf approx. 140 bpm) which turned out to be ventricular became a serious problem. Cardiac function remained stable and troponin and transaminase level normal as well. On day 30 ventricular arrhythmia persisted and did not respond to lidocaine and electroconversion. The patient was neurologically normal for his age, alert and fine. He died in the early morning of the 32 day of life from untreatable arrhythmia. A final echocardiography showed a hypodynamic right ventricle, while the left ventricle was still in a sufficient status (SF about 28-30%). Autopsy revealed a myocardium without necrosis and inflammatory infiltrations. Myocytes with reduced amount of myofibrils were found. In the myocardium of the right heart there was fibrosis. Investigation of the respiratory chain in autopsy samples of the heart showed a moderate decrease of complex I 4.1 mU/mg protein (normal 5.5-51.5 mU/mg protein) and complex IV 64 mU/mg protein (normal 73.2-516.6) in the left ventricle. In the right ventricle the activity of complex I was not detectable and complex IV was reduced to 42 mU/mg protein. Normal activities were found in skeletal muscle and fibroblasts.

Individuals P8-P9 (Family Leeds) were born to consanguinous parents of Pakastani origin and affected children were homozygous for the *PPA2* mutation c.500C>T (p.Pro167Leu), as also seen in Family Basal. Individual P8, a girl, was seemingly well with normal growth and development until the age of 5.5 months. She was then admitted to Bradford Royal Infirmary following a 24 hour history of vomiting and diarrhoea and had suffered a seizure at home. She had further seizures on arrival, and a poor response to treatment so was intubated and ventilated. A CT head scan was normal. Echocardiogram showed poor contractility and a small amount of tricuspid regurgitation. She was transferred to Sheffield Childrens Hospital and had a cardiac arrest on route. She then suffered further multiple cardiac arrests and despite maximal attempts at resuscitation she eventually died during the course of these. On post mortem examination there were no specific macroscopic abnormalities. Infection screen identified rotavirus in the stool. The brain showed hypoxic injury. The liver showed mild fatty change. Skeletal survey, metabolic and toxicology screens were normal. Fatty acid oxidation levels were carried out on skin fibroblasts and were normal. The heart appeared normal in size, shape and structure. Histology of the heart showed areas of recent necrosis, thought to be related to the recent cardiac arrests. There was also evidence of long-standing myocyte loss with increased interstitial collagen and focal myocyte fibre disarray in the left ventricle and interventricular septum. The disarray was considered insufficient for a diagnosis of Hypertrophic Cardiomyopathy. Tests for myocarditis were normal. CSF glucose was low, but this was performed on a post mortem sample.

Individual P9, a girl, suffered a viral illness at the age of 8 months and then suffered a weeks history of increasing hypotonia and weakness. CK was 15,000 at this time and plasma lactate was raised at 5. Free carnitine was normal at 43.6, but propionylcarnitine raised at 2.38. Urine organic acids were normal. There was no involvement of respiratory muscles. Renal function was normal. Over a period of two weeks her weakness and hypotonia improved and her CK reduced. Echocardiogram at this time was normal. She presented again at the age of 11 months to the Leeds General Infirmary with diarrhoea and vomiting, her oral intake was poor and she was not passing urine. She became increasingly drowsy and capillary refill was prolonged at 3 seconds. An initial blood gas showed pH 6.9, bicarbonate 13.8, base excess -13.1, lactate 8.7 and glucose 6.18. She then suffered a focal seizure with lateral gaze to the left and left sided upper limb jerks. This became a generalized seizure which lasted 12 minutes. She was given IV Lorazepam, a fluid bolus, IV antibiotics and acyclovir. Further seizures followed which were treated with IV Lorazepam, Phenytoin and PR Paraldehyde. At three hours following admission seizures had settled but she was still drowsy. CT head scan at this time was normal. Not long after this she suffered a further focal seizure involving the left upper limb. She then suffered a cardiac arrest and was intubated and ventilated. Maximal CPR was continued for 20 minutes but was not successful. Post mortem examination showed very extensive fibrosis of the heart muscle and normal appearance of the skeletal muscle. The brain looked normal at post mortem. Norovirus infection was confirmed on stool samples from admission. Respiratory chain analysis on peripheral muscle tissue was normal, and histological and histochemical assessment of muscle biopsy did not reveal any major mitochondrial abnormalities.

6.2.1 Family N Ireland

Family N Ireland consists of one affected and one healthy sibling born to unrelated healthy parents from Northern Ireland. Individual P10 died at 11 months following diarrhoea, vomiting, focal then generalised seizure and cardiac arrest. Investigation of the respiratory chain in the autopsy sample of the cardiac muscle revealed a decreased activity of complex I (0.026 units/unit citrate synthase, normal 0.125 0.048). In the skeletal muscle tissue, activities were within the normal range. Extensive fibrosis of the heart muscle was observed.

6.3 PPA2 activity

In silico modelling of *PPA2* variants (as reported in the previous chapter), although useful, did not provide direct evidence that the mutations described here in family Christchurch affect PPA2 function causing the condition seen in all four siblings of our index family. Direct biochemical testing of affected tissue can provide valuable information on pathogenicity, however suitable tissue is difficult to obtain, or is of limited quantity and/or stability. Use of model organisms that show conservation with the gene/protein of interest can provide excellent biological evidence of mutation pathogenicity, but are also an investment in terms of both time and money. There are several approaches to using a model organism to examine the effect of a potentially pathogenic mutation. Often creation of a knock-out (KO) strain is appropriate where the gene of interest is interrupted or made non-functional by the introduction of mutations in key regulatory regions. This can provide an insight into the normal gene function, as the developing organism will likely show defects related to the loss of the gene expression. Both of these methodologies have been employed to investigate *PPA2* mutations and the results of these presented below.

6.3.1 Western blotting of PPA2 protein in human tissue

Western blotting was used to determine levels of PPA2 protein in various tissue types from affected patients and was performed by collaborators from Paracelsus Medical University Salzburg. Normal levels of PPA2 protein were observed in fibroblasts obtained from individuals P5-P7. Autopsy muscle tissue of individual P9, however, showed decreased levels of protein, though this was inconsistent with the elevated levels seen in muscle tissue from P6 and P7 who carry the same *PPA2* mutation. In heart autopsy material from P10, we noted decreased PPA2 levels and also of a complex I structural protein (subunit NDUFS4), correlating with the observed decrease in complex I activity in this tissue. In the cardiac autopsy sample of P7, PPA2 and complex I subunit levels were decreased, as well as expression of the mitochondrial marker protein porin. This is suggestive of a more general reduction of mitochondrial number possibly due

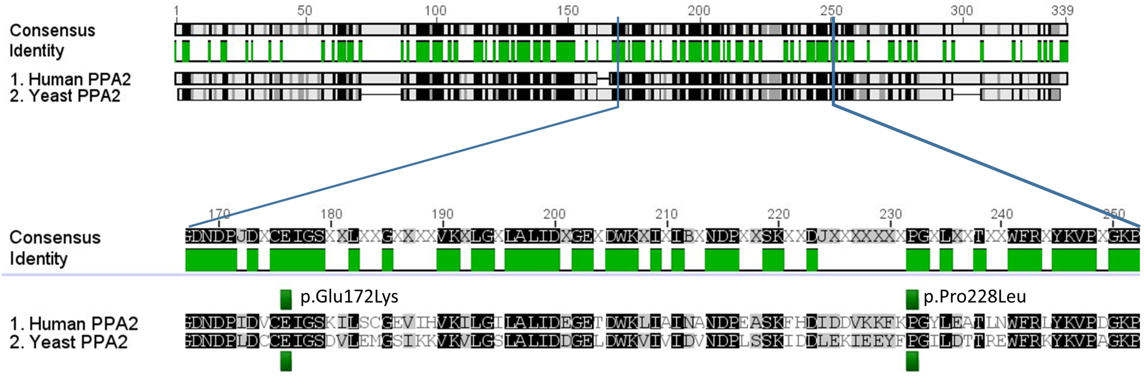

1
50
100
150
200
250
300
339
Consensus
Identity
1. Human PPA2
2. Yeast PPA2
170
180
190
200
210
220
230
240
250
Consensus
Identity
GDNDPJDXCEIGSXXLXXGXXXXVKXLGXLALIDXGEXDWKXIXIBXNDPXXSKXXDJXXXXXXXPGXLXXTXXWFRXYKVPXGKP
p.Glu172Lys
p.Pro228Leu
1. Human PPA2
2. Yeast PPA2
GDNDPIDVCEIGSKILSCGEVIHVKILGILALIDEGETDWKLIAINANDPEASKFHDIDDVKKFKPGYLEATLNWFRLYKVPDGKP
GDNDPLDCCEIGSDVLEMGSIKKVKVLGSLALIDDGELDWKVIVIDVNDPLSSKIDDLEKIEEYFPGILDTTREWFRKYKVPAGKP

Figure 6.2 Human PPA2 and Yeast PPA2 protein alignment
Alignment of homologous proteins showing conserved residues particularly at Glu172 and Pro228.

to changes in tissue composition (Figure 6.3).

6.3.2 *ppa2* knockout growth phenotype on aerobic media

The *S. cerevisiase* pyrophosphatase (*ppa2*) shows significant homology to human PPA2, particularly at key residues of interest (Figure 6.2). Previously *ppa2* knockout yeast strains have been shown to have a growth defect phenotype when plated on an aerobic media where the yeast are required to utilise the mitochondrial respiratory chain [199]. This growth phenotype-was confirmed in the BY4742 S. cerevisiae strain by comparing growth of wild type (WT) and *ppa2* knockout yeast on yeast peptone (YP) glucose media which allows fermentative growth, to growth on YP galactose media which requires oxidative growth. As previously established, growth on galactose was abolished in the *ppa2* knockout yeast strain, indicating that mitochondrial respiration was not available (Figure 6.3).

6.3.3 Oxidative stress sensitivity

Diamide is a substance that oxidises intracellular thiols and mimics oxidative stress in yeast grown on diamide-containing media. Sensitivity of *ppa2* knockout yeast vs WT yeast was assessed at a range of diamide concentrations (0mM, 1mM, 1.25mM, 1.5mM). At 1mM and 1.25mM concentrations the clearest sensitivity is observed in *PPA2* knockout BY4742 *S. cere-*

visiae cells characterised by a reduced growth phenotype, approimately estimated to be in the range of 20-fold more sensitive than WT to oxidative stress (Figure 6.3).

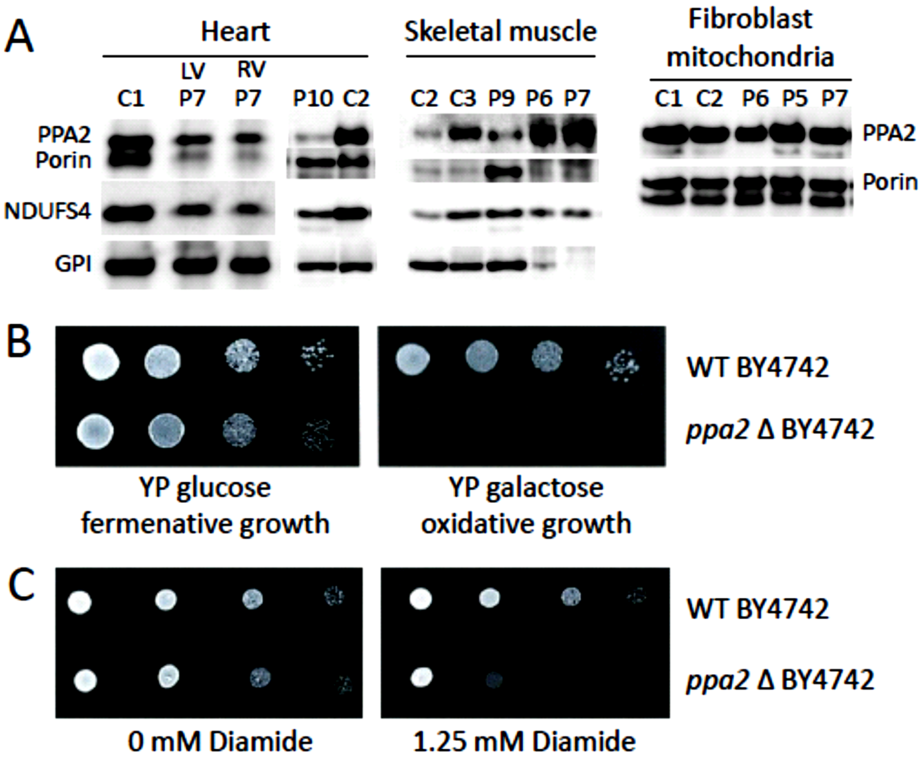


Figure 6.3 Western blot analysis and yeast PPA2 knockout strains
A Western blot analysis in heart, skeletal muscle and fibroblast mitochondria in individuals with *PPA2* mutations. NDUFS4; a subunit of complex I, porin; a mitochondrial outer membrane protein, GPI; a glucose-6-phosphate isomerase, which is a cytosolic housekeeping protein. LV, left ventricle; RV, right ventricle. **B** Growth defect of a *PPA2* knockout (KO) *S. cerevisiae* strain on aerobic medium showing that *PPA2* is critical for mitochondrial respiration. **C** Oxidative stress sensitivity of *PPA2*-KO *S. cerevisiae* as measured by growth defect of diamide, which oxidises intracellular thiols and mimics oxidative stress in yeast.

6.3.4 Pyrophosphatase activity of recombinant PPA2

Pyrophosphatase activity was compared in isolated mitochondria from fibroblasts of controls and affected individuals by measuring the hydrolysis of PPi and quantifying the orthophosphate (Pi). The activity of PPA2 was significantly decreased in isolated fibroblast mitochondria from affected individuals P5 and P7 at all PPi and MgCl2 concentrations investigated (0.001 - 0.1

mol/l and 0.5 or 3.0 mmol/l respectively). Inactivation by CaCl2 was similar in affected individuals compared to controls. Equal amounts of either wild type or mutant recombinant PPA2 proteins were used for the pyrophosphatase activity assay. Compared to wild type the p.Pro167Leu and p.Glu172Lys mutations showed 5-10% residual activity at PPi substrate concentrations 18-500 mol/l. The p.Pro228Leu mutation had a residual activity of 24-28% in this concentration range compared to wild type (Figure 6.4). The activities of wild type and mutants were similarly sensitive to inhibition by Ca^{2+} (data not shown). This work was performed in Austria, by Hans Mayhr's laboratory.

6.4 Discussion

6.4.1 *PPA2* in a mitochondrial context

Production of ATP is a coordinated system of two processes. Electrons (in the form of hydrogen ions derived from NADH and FADH2) are transported through the chain in the following manner. Complex I, complex II, coenzyme Q, complex III, cytochrome c, complex IV, before interacting with molecular oxygen to form water. Simultaneously, protons are pumped across the inner membrane from the matrix to the intermembrane space by complexes I, III and IV. ATP is produced by the movement of these protons back into the matrix via complex V [208]. On the face of it, it would appear that dysfunction of a mitochondrial pyrophosphatase would inhibit ATP production due to reducing the pool of free Pi ions available for uptake into complex V during ATP synthesis. However, the presence of cytosolic Pi molecules that can be imported across the inner membrane complicates this theory. Movement of many substrates across the largely impermeable inner mitochondrial membrane is facilitated by mitochondrial carriers (MCs), part of a super-family of nuclear encoded secondary transport proteins. These carriers have roles in a range of metabolic pathways and cell functions, and mutations in the genes encoding them lead to a number of identified diseases [211]. In humans, over 48 MCs have been identified, with approximately 39 distinct functions. Several classes of transporters are required to facilitate the transport of varying substrates, including nucleotides, carboxylic

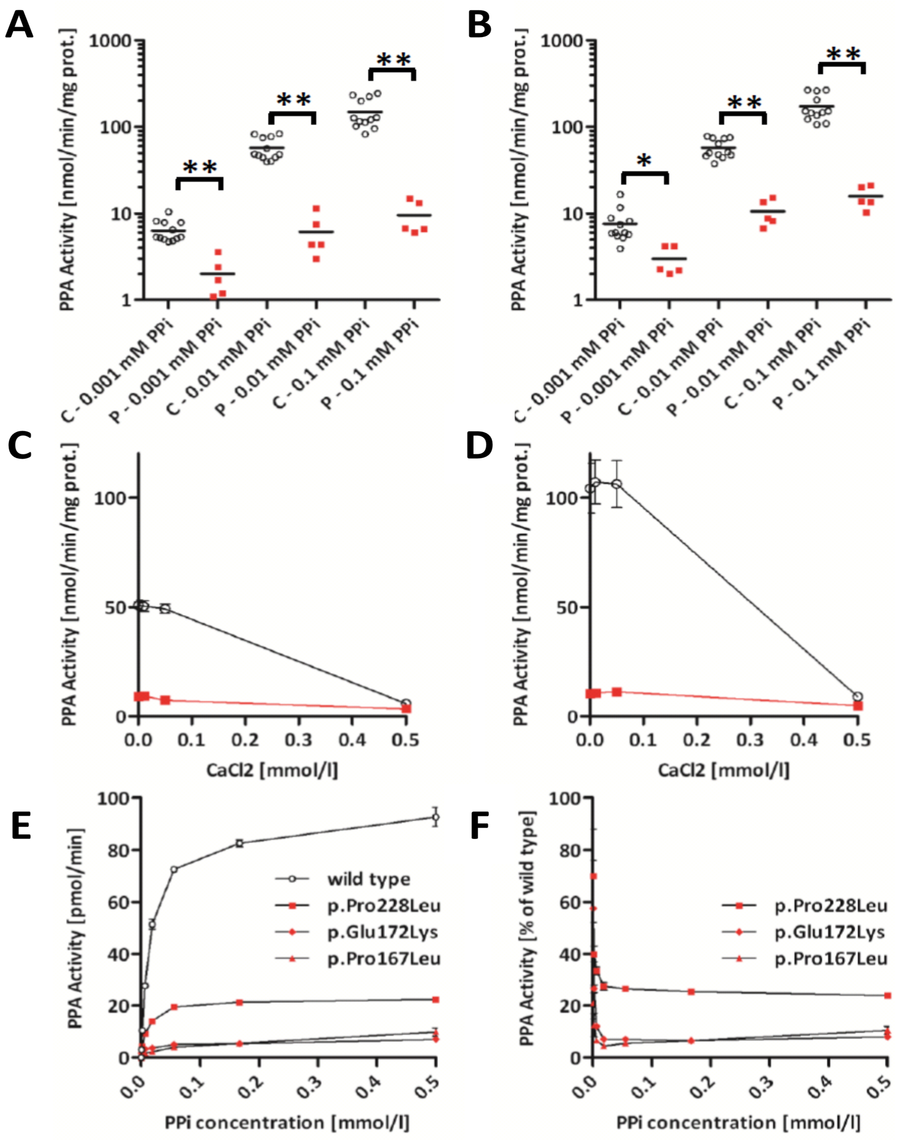


Figure 6.4 PPA2 enzyme activity in fibroblast mitochondria
A - **B** Activity of inorganic pyrophosphatase in fibroblast mitochondria from affected individuals (P) P5 and P7 compared to 12 controls **C** at different PPi concentrations and either **A** 0.5 mmol/l MgCl2 or **B** 3.0 mmol/l MgCl2. **C** - **D** Inhibition of inorganic pyrophosphatase in fibroblast mitochondria from affected individual P5 (red squares) and 3 controls (black circles) at different CaCl2 concentrations and either (**C** 0.01 mmol/l PPi or **D** 0.1 mmol/l PPi. **E** Pyrophosphatase activity of equal amounts of recombinant proteins at different PPi concentrations. **F** Amount of residual activity in different mutants. *P<0.02, **P<0.0002 in Students unpaired t-test.

acids, co-factors, inorganic anions and amino acids [212]. The MCs can be divided into two general functional groups. The first group are involved in metabolic energy generation within the cell, and are associated with the transport of products and substrates required for OXPHOS reactions. This includes the AAC (ADP/ATP carrier), the PiC (phosphate carrier) and the UCP (uncoupling protein) that functions as a proton carrier. The secondary group involves those transporters needed for intermediary metabolism, including, but not limited to, amino acid metabolism and mtDNA replication. Transport of metabolites across the inner mitochondrial membrane is performed in a highly regulated manner, so as to not disrupt the H^+ electrochemical potential gradient across the membrane generated by the OXPHOS chain [213]. Mitochondrial carrier reactions utilise either the electrical or chemical nature (or both) of the H^+ electrochemical potential gradient and/or the solute concentration gradient across the inner mitochondrial membrane. The transport of ADP/ATP by the ACC (encoded by the nuclear gene SLC25A4) is via an antiporter mechanism, where one $ATP4^-$ is exchanged for one $ADP3^+$ molecule simultaneously. ATP preferentially travels along the membrane potential across the inner mitochondrial membrane to the cytosol. The high electrical component of the proton motive force provides the power to eject $ATP4^-$ against the concentration gradient. In contrast, the phosphate carrier (PiC) mechanism of Pi movement is electroneutral co-transportation. Here, Pi^- is translocated with an equivalent amount of H^+ so there is no transfer of charge. This reaction is dependant on the pH gradient across the inner mitochondrial membrane which regulates the distribution of anionic and cationic solutes. The PiC is encoded by SLC25A3, which is alternately transcribed to produce two isoforms PiC-A which is muscle/heart specific, and PiC-B which is ubiquitous and has a higher substrate affinity. The differing affinity of these isoforms is used to control import of Pi and therefore rate of ATP production during muscle contraction [212, 211, 214, 212, 215]. Considering this highly regulated membrane transport of molecules, particularly the PiC which imports Pi for complex V, the simplistic hypothesis that PPA2 dysfunction reduced the Pi available for complex V was disregarded. A defect in PiC-A-mediated uptake has been previously seen in the instance of a homozygous mutation within exon 3A (alternately spliced in PiC-A and PiC-B). The defect led to inhibition of ATP synthesis by OXPHOS, the carrier proteins function appeared severely compromised, and Pi was not being supplied to complex IV. This causes a terminal shortage of energy supply within the heart/muscle cells, but not within all other tissues as they are supplied with Pi by PiC-B. Patients with a PiC deficiency suffer from progressive hypertrophic cardiomyopathy, muscular hypotonia, elevated plasma lactate levels and lactic acidosis which usually leads to heart failure within the first 4 to 9 months of life [212]. This suggests that the levels of Pi provided by hydrolysis of PPi within the mitochondria are likely negligible in terms of supply to complex V and other mechanisms of pathology need to be considered.

6.4.2 Theoretical proposal of *PPA2* disease mechanism

A proposed alternate mechanism for *PPA2*-related mitochondrial dysfunction based on our limited knowledge of PPA2 function is shown in Figure 6.5. If PPA2 activity is reduced due to pathogenic mutations, there are two immediate effects. Firstly, the thermodynamic pull provided by hydrolysis of PPi in many cellular synthesis reactions is likely to be reduced, and may result in a decrease of critical products involved in the mitochondria and wider cell. Secondly, it is expected that PPi will accumulate in the mitochondrial matrix rather than be hydrolysed to Pi. Increased matrix Pi may lead to increased membrane permeability, mitochondrial swelling, and in extreme cases, initiation of the mitochondrial matrix permeability transition (MMPT) which can cause mitochondrial and cell death. The current knowledge of these cellular mechanisms is explored below in the context of PPA2 dysfunction.

Thermodynamic pull for synthesis

The hydrolysis of inorganic pyrophosphate to inorganic phosphate is a strongly exergonic reaction with a G of -33.7kJ/mol, comparable to that of ATP hydrolysis at -35.7 kJ/mol [197]. It is the rapid hydrolysis of PPi which is vital for providing the thermodynamic pull for over thirty cellular reactions (including all individual aminoacyl-tRNA synthetases), detailed in Table 6.1. In the instance of reduced PPA2 function, it is reasonable to assume that the equilibrium of these reactions will be lost with the balance pushed away from synthesis. Pi-producing reactions can be generally divided into functions related to RNA/DNA maintenance, the TCA cycle, and the Electron Transport Chain, all of which are potentially critical in maintaining cell function.

Several synthesis reactions dependant on PPA2 to provide thermodynamic pull are of particular interest in the context of mitochondrial and cardiac disease seen in the pedigrees investigated in this thesis. Bifunctional coenzyme A synthase (CoA synthase) is an enzyme which catalyses the last two steps in CoA synthesis from pantothenic acid (vitamin B5). CoA is an essential requirement for acetyl-CoA synthetases, which are then required for oxidation of pyruvate in the TCA cycle as well as having a role in synthesis and oxidation of fatty acids. PPA2 hydrolysis drives the production of both CoA and Acyl-CoA. The double hit of reduced CoA as well as Acyl-CoA from limited substrate (CoA) and PPA2 dysfunction may result in limited Acyl-CoA entry to the TCA cycle from both glycolysis and fatty acid metabolism. PPA2 dysfunction may also negatively effect reactions directly involved in the ETC such as CoQ synthesis and FAD synthesis. CoQ (ubiquinone) serves as a redox carrier in the ETC and is a lipid-soluble antioxidant. One of the final reactions in the biosynthesis of CoQ is the prenylation of parahydroxybenzoate with an all-trans polyprenyl group. This step is catalysed by CoQ2 which is a product of a *PPA2* driven reaction. In addition, formation of decaprenyl-diphosphate synthase subunits 1 and 2 are driven by PPA2 hydrolysis. This interferes with the formation of side chains of ubiquinone-10, which is a known cause of Q10 deficiency. FAD-dependent proteins function in a large variety of metabolic pathways including electron transport, DNA repair, nucleotide biosynthesis, beta-oxidation of fatty acids, amino acid catabolism, as well as synthesis of other cofactors such as CoA, CoQ and heme groups. During the TCA cycle, succinate dehydrogenase (complex II in the electron transport chain) requires covalently bound FAD to catalyze the oxidation of succinate to fumarate by coupling it with the reduction of ubiquinone to ubiquinol. At least 20 tRNA synthetases are products of a PPA2 driven synthesis reaction, as is tRNA dimethylallytransferase involving charging of amino acids to their cognate tRNA, providing the substrates for global protein synthesis, as well as stabilisation of tRNA structure. In particular, Alanyl-tRNA synthetase (mtAlaRS), encoded by AARS2 has been previously associated with fatal infantile hypertrophic mitochondrial cardiomyopathy, with near-total deficiency of respiratory complexes I, III, and IV in heart tissue [216]. Additionally tRNA dimethylallytransferase (encoded by TRIT1) has been associated with severe combined mitochondrial respiratory chain defects and corresponding perturbation in mitochondrial protein synthesis [217]. The other products affected by PPA2 dysfunction include DNA-directed RNA polymerase, Poly(A) RNA polymerase, Polyribonucleotide nucleotidytransferase (all involved in RNA synthesis, expression, and RNA metabolism), as well as Deoxyuridine 5'-triphosphate nucleotidohydrolase and mitochondrial DNA polymerase catalytic subunit (involved in pyrimidine metabolism and general mitochondrial synthesis). Key effects that may arise due to PPA2 dysfunction may include a reduction of CoA and Acyl CoA which limits the Acyl CoA entry to the TCA cycle from both glycolysis and fatty acid metabolism, in turn providing less electrons to the ETC. A reduction in CoQ may impair the antioxidant effect of CoQ and possible lead to a CoQ deficiency. Reduced tRNA synthesis and impaired mRNA maintenance due to less poly(A) affects the stability of the structure, while inclusion of more dUTP in DNA instead of dTTP (due to dUTP being present in the deoxynucleotide pool in higher concentrations) can also lead to a hyperactive futile cycle of DNA repair. It would be interesting to investigate the possiblity of CoQ suplementation in patients with PPA2 dysfunction.

PPi accumulation in mitochondrial matrix

Although there is not a specific transporter for the movement of PPi, it appears that this substrate can exit the mitochondria via the AAC in exchange for cytosolic ADP. The transport affinity of PPi is extremely low, and the activity of PPi exchange is about 1/10th of that of ADP/ATP exchange [213]. Therefore PPi is inefficiently exported from the matrix. It is also suggested that the AAC can transport small univalent cations when PPi is bound, and that the conformation of the AAC when PPi is bound has an effect on the permeability of the inner mitochondrial membrane to K^+. Swelling of the mitochondria occurs when an anion such as phosphate is taken up to compensate for K^+ uptake and as osmotic imbalance is created. The mitochondria will swell until the K^+/H^+ antiporter is sufficiently stimulated to balance K^+ entry. When the mitochondrial inner membrane permeability increases, the gradient potential maintained by OXPHOS is lost and ATP synthesis is severely is impaired. A further consequence of PPi binding to the AAC will be seen in the exchange of adenine nucleotides into the mitochondria as PPi is effluxed, although the consequences of this are unclear [218]. These two processes, thermodynamics and PPi accumulation, may be considered separately here but

Table 6.1 Cellular reactions requiring PPA2 involvement
Cellular processes and synthesis products which rely on PPA2 function.

Cellular reactions involving pyrophosphatase activity provided by PPA2

Cellular Process	Synthesis products affected	Function
CoA synthesis	Bifunctional coenzyme A synthase (CoA synthase)	CoA synthase is a bi-functional enzyme catalyzing the last two steps of coenzyme A biosynthesis. Biosynthesis of coenzyme A (CoA) from pantothenic acid (vitamin B5) is an essential universal pathway as CoA is required for Acetyl-CoA synthesis for use in the TCA cycle.
Acyl CoA synthesis	Acetyl-coenzyme A synthetase	Acetyl-CoA is required for oxidation of pyruvate. Fatty acids have roles in energy storage, metabolism and as signalling molecules. These essential functions require activation of the fatty acid by acyl-coenzyme A (CoA) synthetases, such as ACSM3, which form an activating thioester linkage between the fatty acid and CoA.
CoQ synthesis	4-hydroxybenzoate polyprenyltransferase	4-hydroxybenzoate polyprenyltransferase is the catalyst for the prenylation of parahydroxybenzoate in the biosynthesis of CoQ (ubiquinone). CoQ serves as a redox carrier in the mitochondrial respiratory chain and is a lipid-soluble antioxidant
	Decaprenyl-diphosphate synthase subunit 1/2	Decaprenyl-diphosphate is the precursor for the side chain of the isoprenoid quinones ubiquinone-10. Defects in this are a cause of coenzyme Q10 deficiency.
FAD synthesis	FAD synthase	FAD-dependent proteins function in a large variety of metabolic pathways including electron transport, DNA repair, nucleotide biosynthesis, beta-oxidation of fatty acids, amino acid catabolism, as well as synthesis of other cofactors such as CoA, CoQ and heme groups. One well-known reaction is part of the citric acid cycle (also known as the TCA or Kreb's cycle); succinate dehydrogenase (complex II in the electron transport chain) requires covalently bound FAD to catalyze the oxidation of succinate to fumarate by coupling it with the reduction of ubiquinone to ubiquinol
tRNA synthesis/ stability	tRNA synthetases	Aminoacyl-tRNA synthetases are essential and ubiquitous 'house-keeping' enzymes responsible for charging amino acids to their cognate tRNAs and providing the substrates for global protein synthesis.
	tRNA dimethylallyltransferase	The primary functions of tRNA methylations are linked to the different steps of protein synthesis, such as the stabilization of tRNA structure, reinforcement of the codon-anticodon interaction, regulation of wobble base pairing, and prevention of frameshift errors.
RNA synthesis	DNA-directed RNA polymerase	This mitochondrial DNA-directed RNA polymerase regulates mitochondrial gene expression as well as providing RNA primers for initiation of replication of the mitochondrial genome.
mRNA expression maintenance	Poly(A) RNA polymerase	Poly(A) RNA polymerase synthesizes the 3' poly(A) tail of mitochondrial transcripts and plays a role in replication-dependent histone mRNA degradation.
RNA metabolism	Polyribonucleotide nucleotidyltransferase 1	PNPT1 is a subunit of the exosome complex, which is involved in 3-prime-to-5-prime exoribonuclease activity for RNA processing and degradation. The enzyme is predominantly located in the mitochondrial intermembrane space and is involved in import of RNA to mitochondria
DNA repair/synthesis	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Deoxyuridine 5'-triphosphate nucleotidohydrolase removes dUTP from the deoxynucleotide pool, which reduces the probability of this base being incorporated into DNA by DNA polymerases. It also produces the dTTP precursor dUMP. Lack or inhibition of dUTPase action leads to harmful perturbations in the nucleotide pool resulting in increased uracil content of DNA that activates a hyperactive futile cycle of DNA repair.
	Mitochondrial DNA polymerase catalytic subunit	Mitochondrial DNA polymerase catalytic subunit function critical for mitochondrial DNA synthesis.

in the context of the cell, overlap and interact. We suggest that *PPA2* mutations are implicated in reduced progression of of several enzyzmatic reactions, the subsequent build-up of PPi in the matrix, and the decrease of ATP produced within the mitochondria due to this, rather than a lack of Pi supplied to complex V as previously considered.

6.4.3 Alcohol sensitivity and *PPA2*

A striking feature in the index family with segregating *PPA2* mutations is an exquisite alcohol sensitivity, which may act as a "trigger" effect on arythmia and heart function. A proposed mechanism for a compounding influence of ethanol ingestion on PPA2 dysfunction is outlined in Figure 6.6. Ethanol metabolism predominantly occurs in the liver, although the effects of impaired metabolism can be wide-ranging as evidenced in the common 'flushing' reaction seen in individuals who accumulate toxic acetaldehyde (a by-product of ethanol metabolism). Ethanol is oxidised to acetaldehyde by alcohol dehydrogenase (ADH) in the cell cytosol, which is then metabolised to acetate by aldehyde dehydrogenase (ALDH2) within the matrix. The resulting acetate has to be activated to acetyl-coenzyme A, which is accompanied by the formation of equimolar amounts of PPi. This esterification reaction is catalyzed by short-chain acyl-CoA synthetases encoded, for example, by ACSS1, an enzyme with high expression in heart mitochondria [219]. Both of these enzymatic reactions involve reducing NAD^+ (nicotinamide: an intermediate carrier of electrons) to NADH. The reducing equivalent of the NADH produced in the cytosol is transported into the mitochondria by molecules known as the malate-aspartate shuttle [220]. An increase in NADH within the cell during ethanol metabolism results in a reduced cytosolic and mitochondrial environment, and cells are then particularly vulnerable to to damage by by-products such as acetaldehyde. The ratio of NADH to NAD^+ determines the rate at which the reactions can occur, therefore rapid ethanol metabolism relies on efficient oxidation of NADH within the mitochondrial electron transport chain. An undesirable outcome of NADH oxidation is the production of reactive oxygen species (ROS). These act by removing hydrogen atoms from other molecules, and converting these molecules into highly reactive free radicals. We have demonstrated that *ppa2* knockout yeast show increased sensitivity to oxidative stress, and also suggested a mechanism for increased sensitivity due to a reduced CoQ antioxidant effect

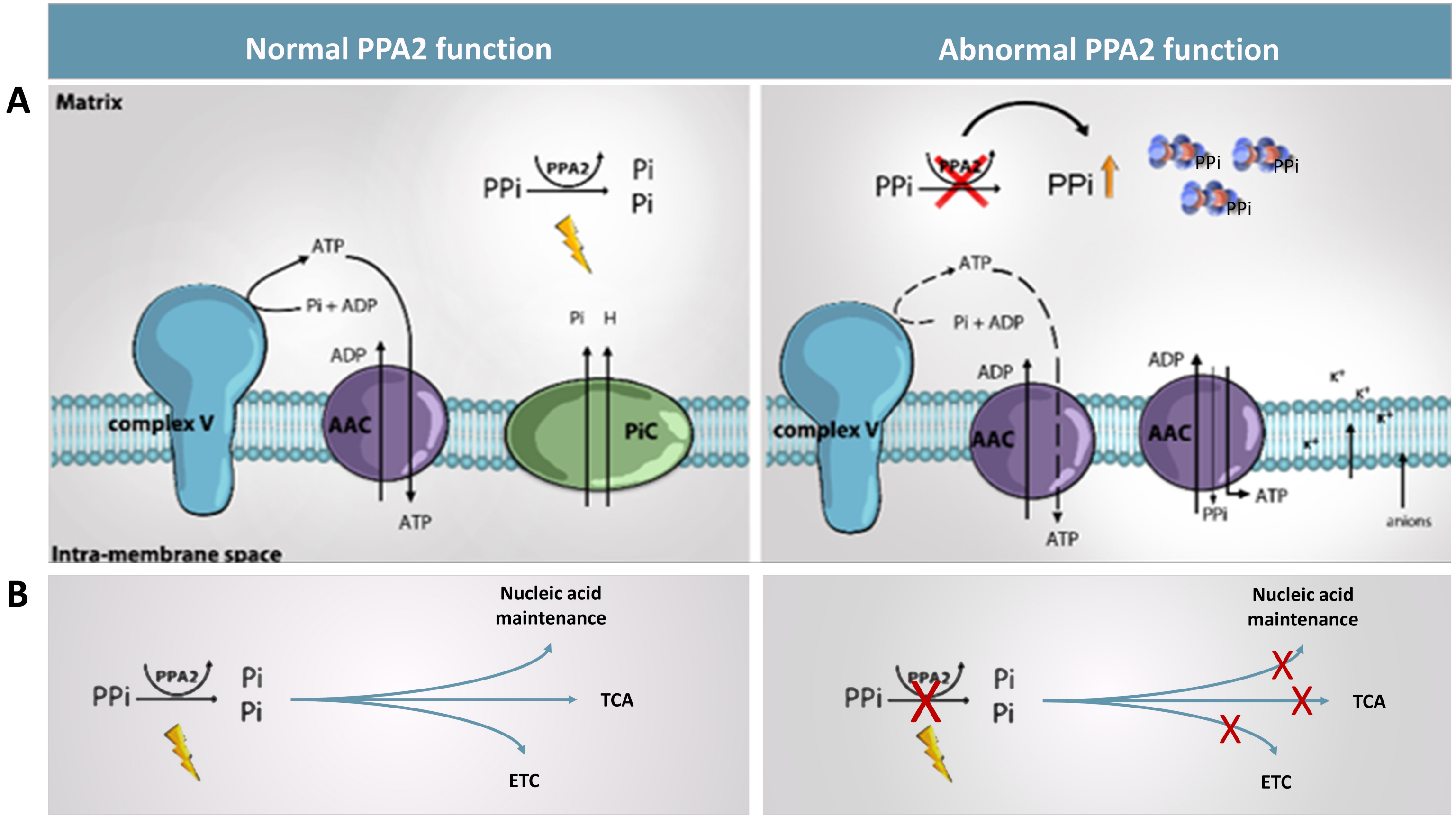


Figure 6.5 Proposed pathogenic mechanisms of PPA2 dysfunction

A Left: Proposed normal function of PPA2 in the mitochondria. Pi is required to enter complex V of the OXPHOS chain as part of ATP synthesis, though Pi can be imported into the matrix by the specialised Pi carrier PiC. ATP production occurs, and ATP is exported from the matrix across the AAC in exchange for ADP. Right: Proposed effect of PPA2 dysfunction. Inefficient hydrolysis of PPi within the mitochondria increases the concentration of matrix PPi. Surplus PPi can be inefficiently exported from the matrix by binding to the AAC and exchanged for ADP in place of ATP. Binding of PPi to the AAC, however, can cause an increase in the inner membrane permeability, particularly to K^+ ions. Once K^+ ions have entered, the matrix will uptake others anions to balance the charge which can lead to swelling of the mitochondria, and in extreme cases, initiation of the mitochondrial matrix permeability transition (MMPT) which can cause mitochondrial and cell death. In the event of either mitochondrial swelling or increased matrix PPi concentration, efficient production and export of ATP may be compromised. **B** Left: PPi is produced as a by-product of many critical reactions, including synthesis of products involved in RNA/DNA maintenance, the TCA cycle, and the Electron Transport Chain. The essentially irreversible hydrolysis of PPi to two Pi molecules provides a thermodynamic pull for these synthesis reactions to proceed. Right: Inefficient hydrolysis of PPi reduces the thermodynamic pull on synthesis reactions resulting in possible reduced synthesis of critical products used by the mitochondria and wider cell.

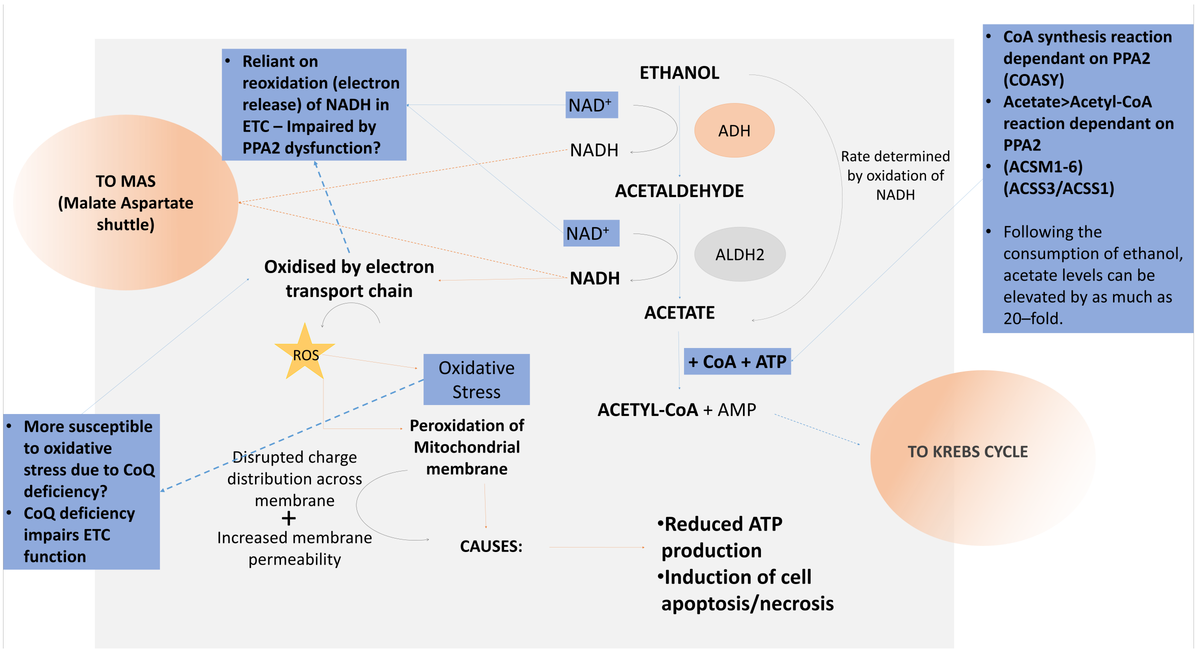

Reliant on reoxidation (electron release) of NADH in ETC – Impaired by PPA2 dysfunction?
ETHANOL
NAD+
NADH
ADH
ACETALDEHYDE
NAD+
ALDH2
NADH
ACETATE
Rate determined by oxidation of NADH
TO MAS (Malate Aspartate shuttle)
Oxidised by electron transport chain
ROS
Oxidative Stress
+ CoA + ATP
ACETYL-CoA + AMP
CoA synthesis reaction dependant on PPA2 (COASY)
Acetate>Acetyl-CoA reaction dependant on PPA2
(ACSM1-6) (ACSS3/ACSS1)
Following the consumption of ethanol, acetate levels can be elevated by as much as 20–fold.
TO KREBS CYCLE
More susceptible to oxidative stress due to CoQ deficiency?
CoQ deficiency impairs ETC function
Peroxidation of Mitochondrial membrane
Disrupted charge distribution across membrane
+
Increased membrane permeability
CAUSES:
•Reduced ATP production
•Induction of cell apoptosis/necrosis

Figure 6.6 Proposed influence of ethanol ingestion on PPA2 dysfunction
A proposed mechanism for a compounding influence of ethanol ingestion. Ethanol metabolism predominantly occurs in the liver, although the effects of impaired metabolism can be wide-ranging as evidenced in the common 'flushing' reaction seen in individuals who accumulate toxic acetaldehyde (a by-product of ethanol metabolism). Ethanol is oxidised to acetaldehyde by alcohol dehydrogenase (ADH), which is then metabolised to acetate by aldehyde dehydrogenase (ALDH2). Both of these enzymatic reactions involve reducing NAD^+ (nicotinamide: an intermediate carrier of electrons) to NADH. An increase in NADH within the cell during ethanol metabolism results in a reduced cytosolic and mitochondrial environment, and cells are then particularly vulnerable to to damage by by-products such as acetaldehyde. The ratio of NADH to NAD^+ determines the rate at which the reactions can occur, therefore rapid ethanol metabolism relies on efficient oxidation of NADH within the mitochondrial electron transport chain. An undesirable outcome of NADH oxidation is the production of reactive oxygen species (ROS). These act by removing hydrogen atoms from other molecules, and converting these molecules into highly reactive free radicals. An excess of ROS results in a state of oxidative stress, and can lead to lipid peroxidation, in particular, peroxidation of the inner mitochondrial membrane. Peroxidation of the membrane has several detrimental effects. Peroxidation of the membrane has several detrimental effects, namely increased membrane permeability, and a reduced electrical charge gradient. These two changes will contribute to a flux of molecules exiting the mitochondria which can induce apoptosis, and reduced ATP production causing necrosis respectively.

(Figure 6.3). An excess of ROS creates a cellular state of oxidative stress, and can lead to lipid peroxidation, in particular, peroxidation of the inner mitochondrial membrane. Peroxidation of the membrane has several detrimental effects, including increased membrane permeability, and a reduced electrical charge gradient. These two consequences will contribute to a flux of molecules leaving the mitochondria which can induce apoptosis, and reduce ATP production leading to necrosis. Affected individuals in our index family reported experiencing generalised myopathy and discomfort after ethanol ingestion. The surviving male (P2) also reported a sensitivity to cooked shellfish (mussels) in vinegar when questioned about a possible effect of acetate. As acetate is formed during ethanol metabolism and it's conversion to acetyl-CoA requires PPA2 hydrolysis, it was of interest to know that some acetate sensitivity was present without prior ethanol intake. The other surviving female (P4) did not recall any sensitivity to vinegar or other foods containing acetate so the response of P2 is likely to be less acute than the exquisite alcohol sensitivity described in this family. Consumption of acetate in these individuals rather than ethanol would result in a build up of acetate due to inefficient conversion to acetyl-CoA, but not acetaldehyde or ethanol which would occur with ethanol ingestion. This reaction does not involve reducing NAD^+ and therefore would not be expected to cause the same effect of increased NADH as ethanol metabolism, perhaps explaining a milder phenotypic effect. In this family, the fibrosis in the heart (likely caused by chronic low-level ATP depletion) may have primed the affected individuals to a secondary trigger event of arrhythmia. A trigger of sudden death is not a new concept, particularly in the case of Long QT syndrome which is known to be associated with genotype-specific triggers of life-threatening arrhythmias. Immersion of the face in cold water in cases of swimming or diving is a common trigger of arrhythmia in patients with Long QT syndrome 1 [221, 222]. In Long QT syndrome 2, auditory stimulus (usually during sleep) such as an alarm or bell has been described to cause syncope and subsequent death [223, 222]. Although no other family had a recorded sensitivity such as this alcohol intolerance, presumably from lack of exposure due to their neonatal stage of life, it is worth noting that both affected individuals in Family Leeds had a history of vomiting, diarrhoea and seizures prior to admission to hospital. Viral infection (rotavirus (first) and norovirus (second)) was confirmed in stool samples taken at time of admission. A norovirus infection was also detected in P10 from the United Kingdom. Although no viral illness was recorded in other individuals, similar symptoms also preceded hospital admission in two siblings of Family Basel. This may indicate that a viral stressor was responsible for adversely affecting mitochondrial metabolism in these cases, in the same way that alcohol was a trigger for arrhythmia in Family Christchurch.

6.4.4 Proposed Mechanism for SCD due to *PPA2* mutations

In addition to considering a mechanism for disease in the index family, we also propose several interactions which may lead to the multiple symptoms which presented in all four families currently identified with *PPA2* mutations. Figure 6.7 illustrates some of the possible links discussed in this thesis into one general scheme, including features such as lactic acidosis and arrhythmia observed in the affected infants with *PPA2* mutations in our other pedigrees. Generally we propose that PPA2 dysfunction due to compounding or homozygous mutations directly impacts the mitochondrial ATP output and increases mitochondrial sensitivity to stress. This occurs through the loss of thermodynamic pull in critical enzymatic reactions, as well as accumulation of PPi in the matrix, as discussed in section 6.4.3. The immediate effects of severe ATP depletion are likely to be lactic acidosis and arrhythmia as glycolysis is employed to provide immediately available energy to the cell. More chronic effects of PPi build-up in the matrix may include mitochondrial swelling and a reduction in membrane potential, as well as a more mild reduction of ATP leading to cardiac fibrosis. The effects of impairing synthesis reactions dependant on PPA2 are largely to do with sensitisation of the mitochondria. Inhibiting CoA and acyl-CoA synthesis can lead to impaired metabolism of alcohol, while reducing Co-Q synthesis impacts the antioxidant protection. These pathways all lead to reduced ATP production, an increased sensitivity to oxidative stress, and reduced Ca^{2+} regulation. In the three additional families identified it is expected that severe ATP depletion had an acute effect causing death from lactic acidosis and cardiomyopathy prior to any chronic damage developing. A feature of partial PPA2 function provided by the assumed less severe Pro228Leu mutation in the index family may be that more chronic effects such as accumulation of cardiac fibrosis occur instead of the acute symptoms expected from near total PPA2 dysfunction. In Family Christchurch ethanol ingestion has been proposed as the trigger underlying sudden death. We propose that the metabolism of alcohol (in mitochondria with limited CoA and acyl-Coa synthesis) significantly increases NADH concentration, and the oxidation of this results in ROS production. In mitochondria already predisposed to oxidative stress (due to reduced CoQ antioxidant effect) this ROS production may precipitate a cascade of reactions that involves peroxidation of the mitochondrial membrane, swelling, Ca^{2+} overload, and opening of the mitochondrial permeability transition pore (MPTP). The MPTP is described as the reversible opening of a pore in the inner mitochondrial membrane that allowed the passage of any molecule of <1500 Da which is observed in response to certain pathological conditions such a reperfusion injury. The result of opening the MPTP will be a loss of proton-motive force, rendering the mitochondria incapable of ATP production and initiating cell death mechanisms. This process can quickly become non-reversible, as once a single MPTP has opened, the mitochondrion is sensitised to Ca^{2+} concentrations which induces multiple pore openings in the same organelle. The Ca^{2+} released upon swelling of this mitochondrion, is subsequently taken up by surrounding mitochondria, which induces further MPTP opening, creating a cascade of organelle failure. MPTP opening is promoted by Ca^{2+}, oxidative stress, high phosphate concentrations and depletion of matrix adenine nucleotides. Opening is, however, inhibited by low pH, which occurs as a result of glycolosis. In the case of reperfusion injury, ischemia causes the respiratory chain to stop producing ATP. When oxygen is reperfused and the respiratory chain restarts, the mitochondria can take up the Ca^{2+} which has accumulated during ischaemia, but also produces a surge of ROS. The combination of oxidative stress and Ca^{2+} concentration, prime the MPTP opening, however, it is only when the intracellular pH returns to normal that MPTP opening can occur. We propose that a similar stress occurs upon consumption of alcohol in the index family. Non-fatal events which led to pain and discomfort may have been a product of lactic acid build up (lactic burn), while the sudden cardiac death events were a result of secondary energy failure of the heart, subsequent to the alcohol-induced effects leading to MPTP opening [224].

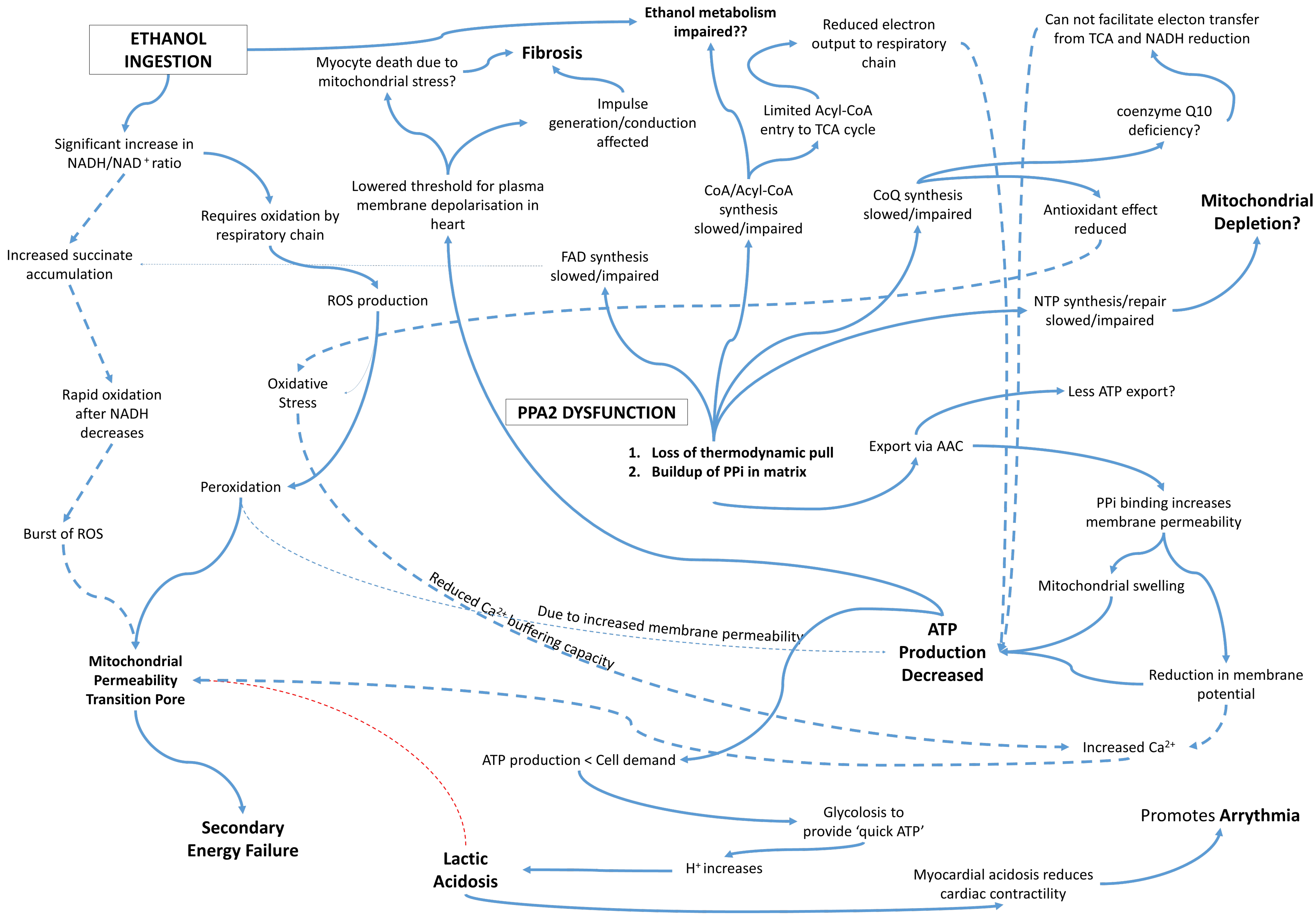

ETHANOL INGESTION
Ethanol metabolism impaired??
Myocyte death due to mitochondrial stress?
Fibrosis
Impulse generation/conduction affected
Reduced electron output to respiratory chain
Limited Acyl-CoA entry to TCA cycle
Can not facilitate electon transfer from TCA and NADH reduction
coenzyme Q10 deficiency?
Significant increase in NADH/NAD⁺ ratio
Requires oxidation by respiratory chain
Lowered threshold for plasma membrane depolarisation in heart
CoA/Acyl-CoA synthesis slowed/impaired
CoQ synthesis slowed/impaired
Antioxidant effect reduced
Mitochondrial Depletion?
Increased succinate accumulation
FAD synthesis slowed/impaired
ROS production
NTP synthesis/repair slowed/impaired
Oxidative Stress
Rapid oxidation after NADH decreases
PPA2 DYSFUNCTION
Less ATP export?
1. Loss of thermodynamic pull
2. Buildup of PPi in matrix
Export via AAC
Peroxidation
PPi binding increases membrane permeability
Burst of ROS
Mitochondrial swelling
Reduced Ca²⁺ buffering capacity
Due to increased membrane permeability
ATP Production Decreased
Mitochondrial Permeability Transition Pore
Reduction in membrane potential
ATP production < Cell demand
Increased Ca²⁺
Glycolosis to provide 'quick ATP'
Promotes Arrythmia
Secondary Energy Failure
Lactic Acidosis
H⁺ increases
Myocardial acidosis reduces cardiac contractility

Figure 6.7 Proposed mechanism for PPA2-related mitochondrial disease
This proposed mechanism incorporates symptoms such as arrythmia, lactic acidosis and cardiac fibrosis seen in patients from affected families, and aims to show how these may be induced by PPA2 dysfunction.

6.4.5 Conclusion

In our index family, we suggest that *PPA2* mutations lead to chronic ATP depletion and sensitisation of the mitochondria (particularly in high energy tissues such as the heart), which leads to the accumulation cardiac fibrosis observed in affected individuals. The symptoms of pain and sudden cardiac death after alcohol ingestion may be attributed to the cumulative effect of this chronic ATP depletion/mitochondrial sensitisation, and then the additional stess to already damaged cardiac tissue as a result of ethanol metabolism. This theory is supported by the observation that symptoms of alcohol-related pain did not occur immediately, but 30-60 minutes after ingestion, and both deaths occurred in the hours following alcohol consumption despite cardiac fibrosis being present prior to death. Our proposed mechanism is based on published biochemistry, but has not been explored experimentally and should be considered only as a starting point for further investigations into the role of *PPA2* in mitochondrial disease. *PPA2* mutations were also identified in three additional families, harbouring mutations which were associated with a more severe phenotype. Analysis of pyrophosphatase activity associated with the various mutations confirmed that less activity associated with a more severe phenotype and earlier death, although variation was observed between identical genotypes. In conclusion, we have identified biallelic missense mutations in *PPA2* as a new cause of mitochondrial cardiomyopathy and sudden cardiac death. This finding highlights a critical role of *PPA2* in mitochondrial function, and warrants further functional investigation. Importantly, mild mutations in *PPA2* may not have an immediate life threatening effect until triggered by a stressor such as viral illness, alcohol metabolism or possibly exercise, predisposing otherwise healthy individuals to sudden cardiac death. Considering the relatively high frequency of *PPA2* mutations present in the ExAC database, it is important that mutations in *PPA2* are screened in clinically relevant populations where cardiac MRI might reveal fibrosis. Moreover, application of an implantable cardioverter defibrillator may prevent sudden cardiac death in at-risk individuals who harbour biallelic *PPA2* mutations.

Chapter 7

Zebrafish modelling of PPA2 dysfunction and mitochondrial disease

7.1 Introduction

Although modelling of ppa2 dysfunction in yeast provided some evidence of pathogenicity, there are limits to what can be extrapolated to a human disorder. For example, when investigating a disorder involving the heart, yeast becomes a far less useful model organism. During vertebrate embryogenesis the heart is the first organ to form and function. In zebrafish (*Danio rerio*), which are transparent and rapidly develop, the heart starts beating as early as 24 hours post fertilisation (hpf), and strong circulation can be seen by 1.5 days post fertilisation (dpf). Despite significant anatomical differences, zebrafish and human hearts are surprisingly similar in terms of beating rates and eletrophysiology. The fish heart is initially formed from migration of myocardial and endothelial cells into a linear tube with a distinct outer myocardial and inner endocardial layer. The heart tube then folds into a looped organ with four distinct chambers; sinus venosus, atrium, ventricle and bulbus artreiosus (outflow tract). Due to their size, zebrafish embryos can obtain enough oxygen through passive diffusion to allow for relatively normal development for several days even in the absence of blood circulation and a functional cardiovascular system. This makes them an good model for the study of cardiac phenotypes, particularly as their transparency allows rapid and reliable *in vivo* visualisation and characterisation of embryos by light microscopy. This makes it possible to study genetic defects that may lead to early lethality in other animals [225].

A useful technique to transiently knock-down a gene of interest in zebrafish (and other models) without needing to produce a stable gene knock-out strain is to use morpholino antisense oligomers (MO). A morpholino is a synthetic oligo which when injected into a developing zebrafish embryo, will target a complementary sequence in a gene of interest and prevent production of protein. Although expression of the gene will not be entirely suppressed, the morpholino enables researchers to view the effect of partial knockdown, particularly useful in cases where a knockout organism would not be viable. Co-injection of morpholino and mRNA is often used in RNA rescue experiments to show that a phenotype in a morpholino injected fish can be 'rescued' by co injection of the WT mRNA which enables near-to normal levels of protein expression.

Zebrafish were successfully used to demonstrate that the ciliopathies Meckel-Gruber syndrome (MKS) and Bardet-Biedl syndrome (BBS), are allelic forms of the same molecular spectrum despite being clinically distinct, and that particular MKS1 mutations were functionally pathogenic. Leitch et al [226] used a combination of morpholino based suppression of mks1, and co-injection with wildtype human MKS1 mRNA to rescue the morphant phenotype, showing variable functionality of previously *in silico* predicted pathogenic variants.

This chapter introduces ongoing research using zebrafish to model PPA2 dysfunction, aiming to produce a stable *ppa2* knockout strain in which to observe the effects of specific *ppa2* mutations seen in the family presented in Chapter 5. This work was performed in collaboration with the Otago Zebrafish Facility, led by Assoc. Prof Julia Horsfield.

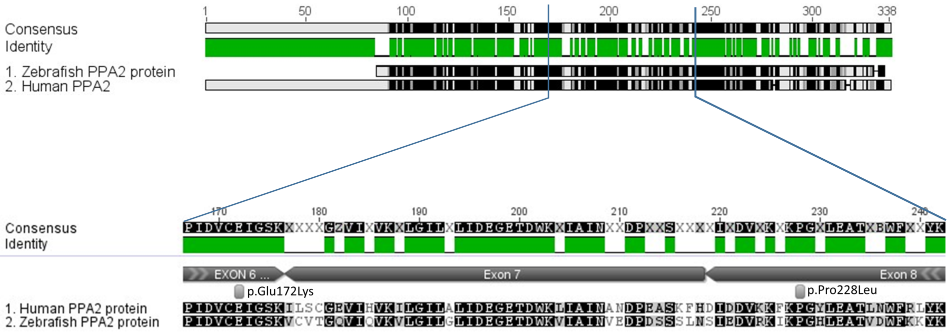


Figure 7.1 Human PPA2 and Zebrafish PPA2 protein alignment
Alignment of homologous zebrafish and human pytophosphatase proteins showing conserved residues, particularly at residue 172 and 228 (human), representing the mutated residues seen in Family Christchurch.

7.2 Zebrafish *ppa2*

Sequence information on the zebrafish *ppa2* gene, on chromosome 1, was obtained from NCBI (Gene ID: 402961). The zebrafish *ppa2* transcript is 1468 base pairs long and encodes 250 amino acids. Comparison to known sequences identified *ppa2* as part of the Pyrophosphatase Superfamily, with 57% identity to the human protein (NCBI BLAST; Human *PPA2* accession number AAH57219.1). The alignment of human and zebrafish PPA2 is shown in Figure 7.1.

7.3 Morpholino knockdown - preliminary experiments

Morpholinos were designed to target two complementary zebrafish *ppa2* sequences: one at the 5'UTR and the other at an internal splice site (Figure 7.2). In addition, a random sequence oligo (MIS-5) was used to act as an injection control, and ensure that phenotypic effects observed were due to *ppa2* expression knockdown and not off target effects of the morpholino.

Titrations of injection volume for the 5UTR MO, SPLX MO, and MIS-5 MO were performed to identify an optimal injection concentration for phenotype characterisation. Too high a concentration results in lethality, while a low concentration may not induce any mutant characteristics in the fish. Optimal concentration was measured by survival and characterisation of phenotype in 1dpf and 2dpf zebrafish embryos. The survival rates of embryos with a 5'UTR or

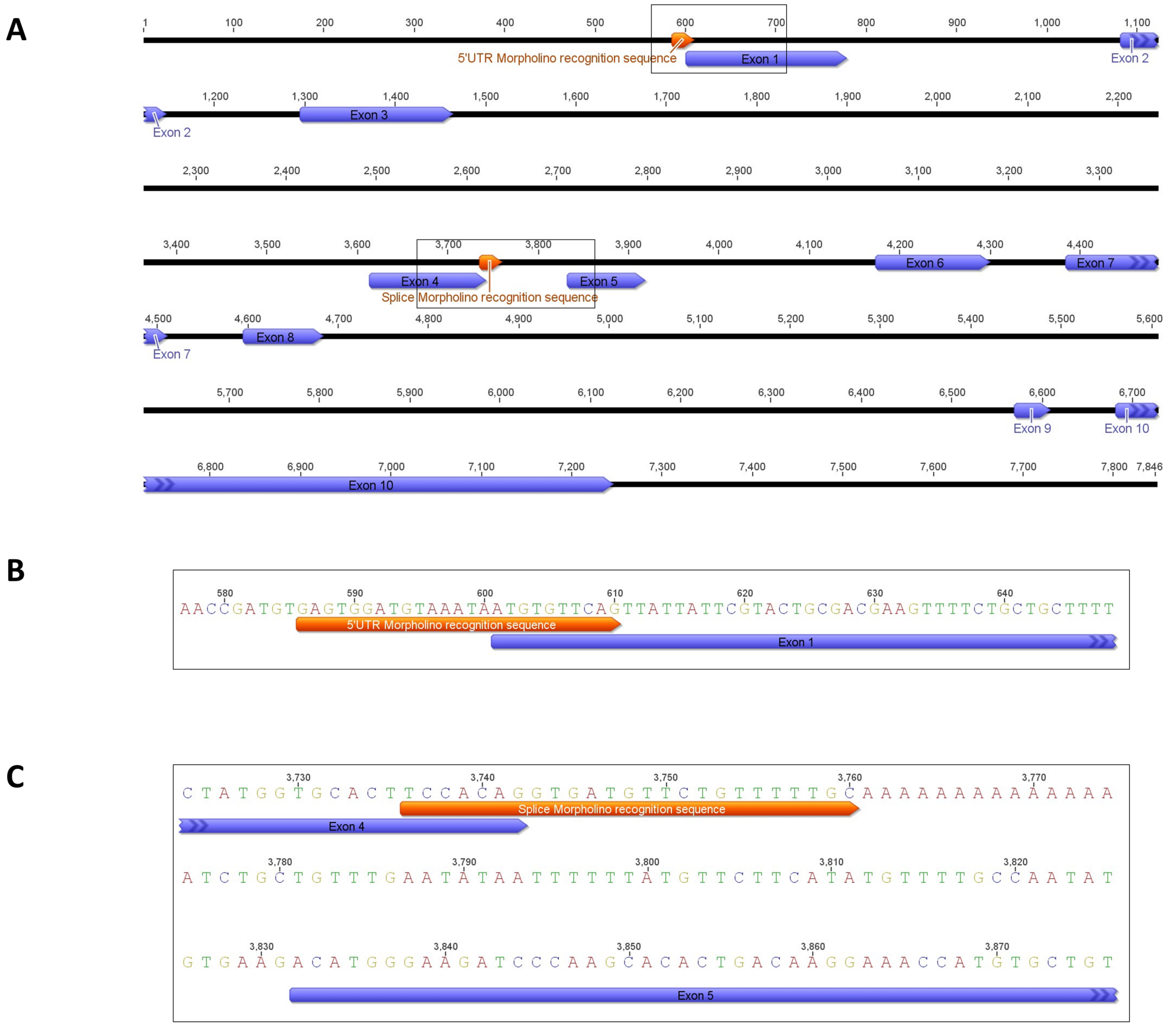


Figure 7.2 Zebrafish ppa2 sequence, showing binding sites of Morpholinos.
A Full zebrafish *ppa2* cDNA sequence showing binding of the 5'UTR MO and Splice site (SPX) MO. **B** Specific sequence details of the binding site for the 5'UTR MO. **C** Specific sequence details of the binding site for the SPX MO.

SPX MO injection concentration of 0.125pmol, 0.25pmol and 0.5pmol at 1dpf was variable, but didn't correlate with a dose dependant effect, which would have suggested any concentration-dependant lethality. Unfortunately the MIS-5 MO appeared to have a significantly detrimental effect and survival rates were too low for use as an suitable control. Injection of sterile water into embryos was used as an injection control in the following experiments. At 2dpf, the concentrations 0.25pmol and 0.5pmol of MO, produced consistent yolk edema phenotypes in the SPLX MO embryos, and a tail malformation of varying severity in both SPLX MO and 5' UTR MO embryos (Figure 7.3). Additionally, some SPLX MO embryos were missing pectoral fins at this stage. A dose-dependant hypopigmentation effect was observed in the UTR MO embryos, but count data was not undertaken for pectoral fins and hypopigmentation effects.

At 5dpf, additional consistent phenotype abnormalities were noted. At a gross level these consisted of protruding jaws and a curved tail tip (Figure 7.4). Based on these initial observations, it was decided that an intermediate titration of MO (0.35pmol) would be suitable for future phenotype experiments to ensure acceptable survival rates and a measurable phenotype was observed.

7.4 RNA Rescue with wildtype *ppa2*

Preliminary results for RNA rescue of the observed *ppa2* knockdown phenotype were achieved by co-injecting embryos with MO and wildtype zebrafish *ppa2* mRNA. MO concentration was 0.35pmol, while WT mRNA was injected at either 50pg or 100pg. One general and one specific feature observed in earlier MO titration experiments were used to quantify the success of the rescue.

A gross severity score (No phenotype/Mild/Medium/Severe) was assigned to individual fish in each treatment group based on several key characteristics observed in preliminary experiments. No phenotype was defined as fish with no obvious defects, and grossly normal development. A mild score designated fish that had a slightly curved or small tail OR yolk adema or cardiac

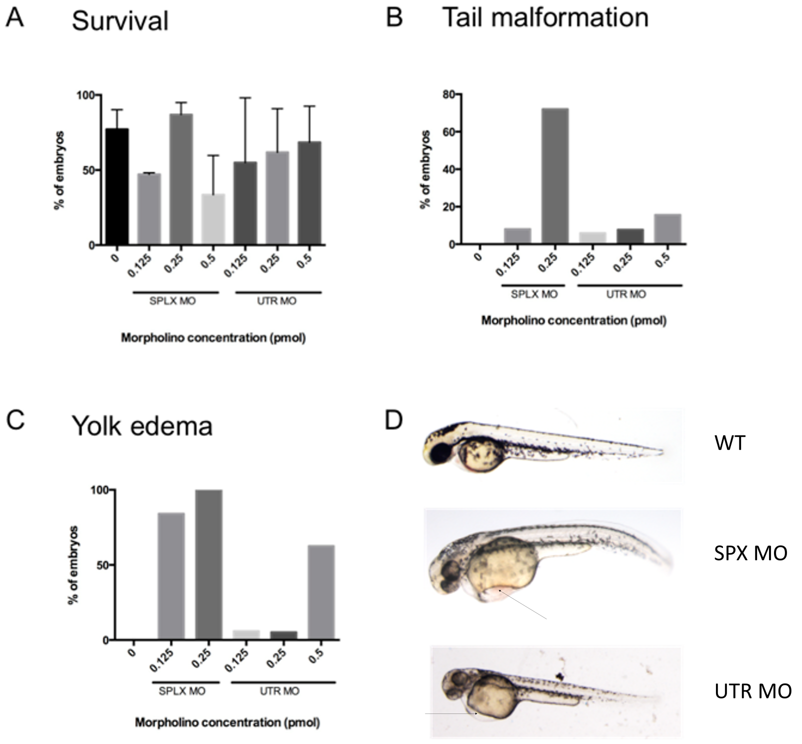


Figure 7.3 Preliminary data from 1dpf and 2dpf zebrafish embryos injected with differing concentrations of both 5'UTR and SPX morpholino preparations (248 embryos counted across 4 experiments) **A** Survival of zebrafish embryos at 2dpf. **B** Observation of tail malformation phenotype. **C** Observation yolk adema phenotype. **D** Examples of 2dpf zebrafish embryo phenotypes in each treatment condition.

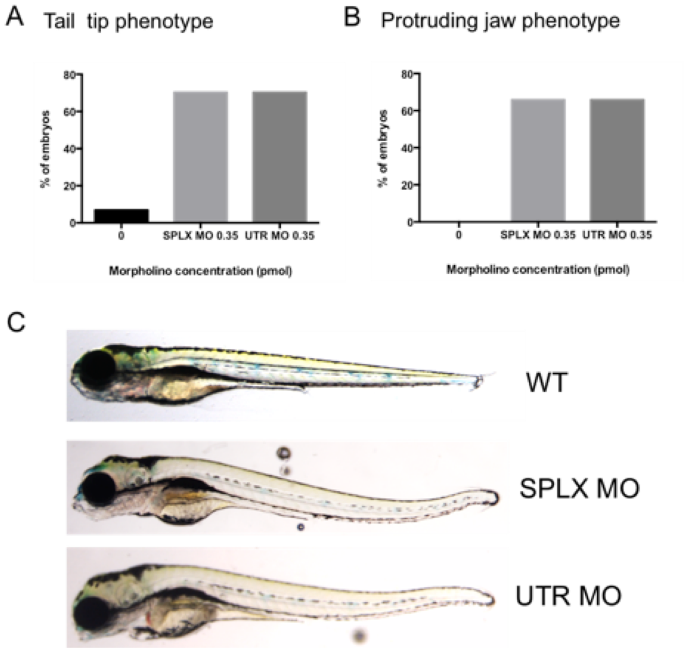


Figure 7.4 Preliminary data from 5dpf treated and untreated zebrafish embryos (92 embryos counted across 2 experiments)
A Observation of tail tip phenotype. **B** Observation of protruding jaw phenotype. **C** Examples of 5dpf zebrafish embryo phenotypes in each treatment condition.

adema/blood backflow or pooling. A medium score required fish to have a slightly curved or small tail, AND yolk adema or cardiac adema or blood backflow or pooling. Severe fish displayed a very curved tail AND yolk adema/cardiac adema/blood backflow or pooling (Figure 7.5). In the 'No treatment' control group, 100% of fish were phenotypically normal. When morpholino was introduced, greater than 68% of fish developed defects scored as either mild, medium or severe. The percentage of fish counted as severe was 25.49%. When mRNA was co-injected with the morpholino treatment, the percentage of fish classed as severe dropped significantly to 7.36% (Note that the p-value displayed relates to the reduction in severe phenotype only, it is based on the Fisher's exact test and only appropriate for comparing one class). The reduction of fish classed as severe (and consequent increase in mild phenotype fish) is consistent with a partial RNA rescue

The second measure used was of abnormal fin development of fish in each treatment group. It was observed in preliminary experiments that PPA2 knockdown fish displayed very stunted dorsal fin growth compared to wildtype fish (Figure 7.6). At 3dpf there was a clear difference in

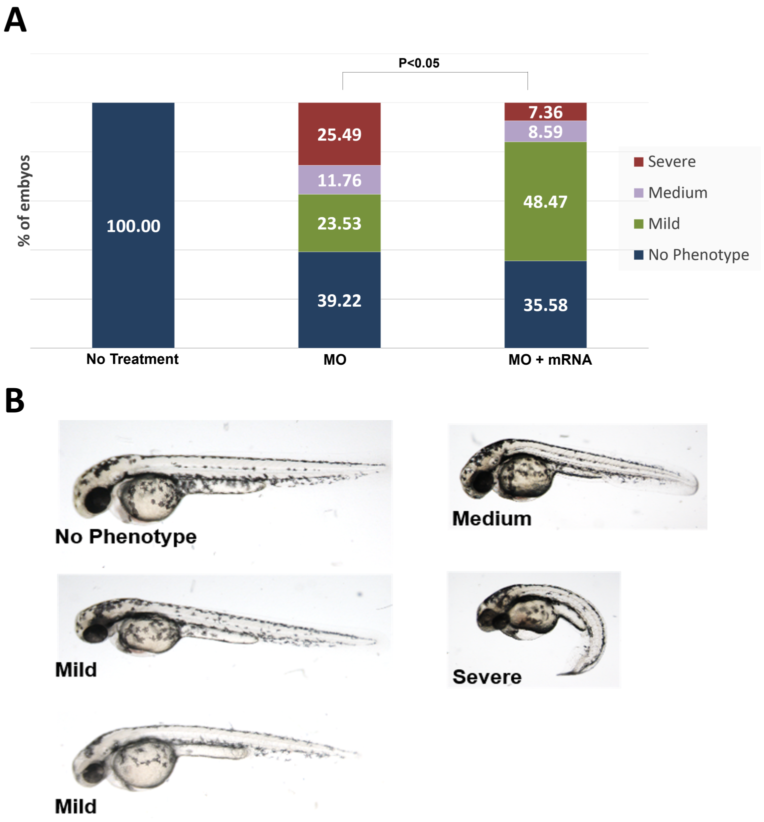


Figure 7.5 RNA rescue of ppa2 knockdown embryos with RNA (50pg) (302 embryos counted across 4 experiments)
A Percentage of embryos displaying a phentoype in four severity groups by treatment type. No phenotype = no obvious defects, and gross normal development; Mild = slightly curved/small tail OR yolk adema/caridac adema/blood backflow or pooling; Medium = slightly curved/small tail AND yolk adema/cardiac adema/blood backflow or pooling; Severe = very curved tail AND yolk adema/cardiac adema/blood backflow or pooling. Fisher's exact test p-value only relates to severe embryos **B** Example phenotypes at 2 dpf indicitive of each grouping in **A**.

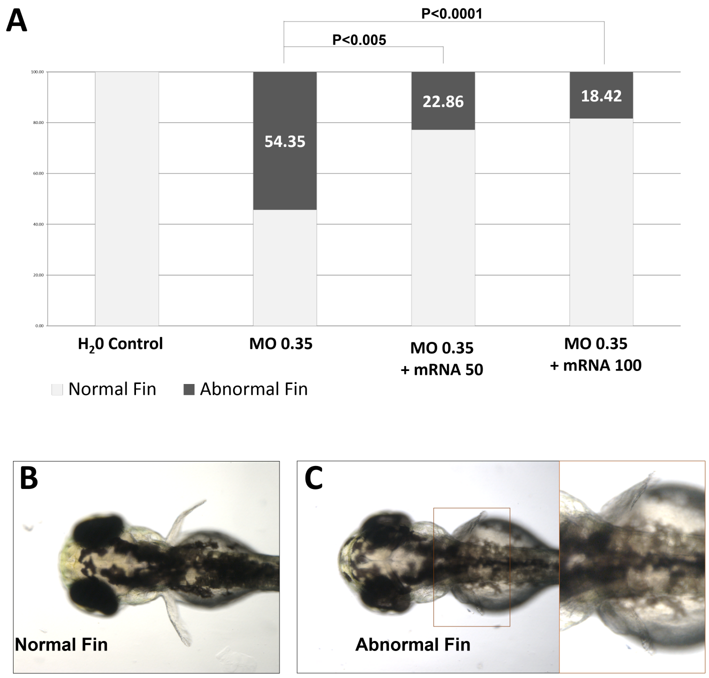


Figure 7.6 RNA rescue of ppa2 knockdown (758 embryos counted across 5 experiments) **A**Percentage of abnormal fins per treatment type. p-value calculated using Fisher's exact test. **B** Example normal fin embryo phenotype at 2 dpf. **C** Example abnormal fin embryo phenotype at 3 dpf (enlarged right)

fin size between knockdown and rescue fish, although dissection to directly measure fin area was not possible. Instead a count of the percentage of fish in each treatment type with abnormal fins was used to demonstrate a significant effect of the RNA rescue group over morpholino only. This was consistent for both a medium dose (50pg) and high dose (100pg) of mRNA with a standard concentration of MO (at 0.35pmol).

7.5 Additional phenotype characterisation methods

Several other phenotypic features were observed as zebrafish embryos developed, and although not enough replicated experiments have yet been performed to accurately quantify these features in treatment groups, various methods of visualising and measuring these features have been trialled.

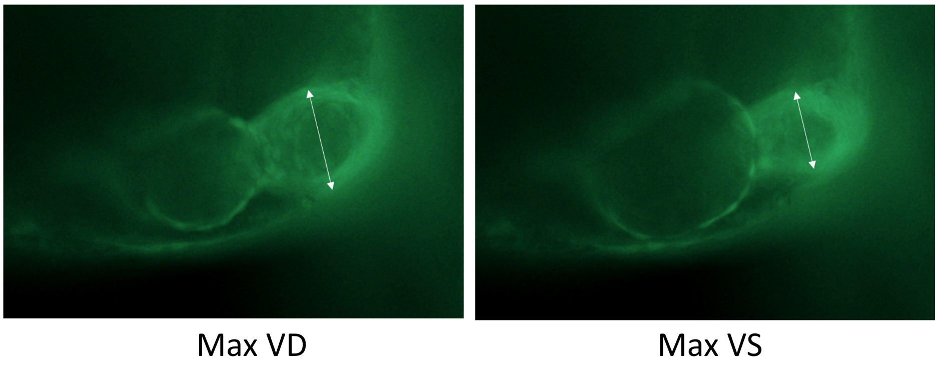


Figure 7.7 Zebrafish ventrical shortening fraction can be used as a measure of heart function The shortening fraction is measured by using the formula: (max VS/max VD)*100 = Shortening faction. Maximum VS and VD are measured in still captures from a movie of a beating zebrafish heart.

7.5.1 Heart related features

Measurement of shortening fraction

The normal function of the heart requires the ventricals to contract and relax in coordinated motion to pump blood through the organ. The ventricular shortening fraction measures the diastolic dimension that is lost during systole and is a measure of heart function. Still images were captured from video clips of 2dfp zebrafish hearts at maximum ventricular systole (VS) and maximum ventricular diastole (VD). An example of maximum VS and VD is shown in Figure 7.7. No significant difference in shortening fraction was observed in preliminary data available. MO injected zebrafish embryos had an average shortening fraction of 83.7% compared to 80.9% in control fish (data not shown).

Measurement of heartrate

Preliminary data of a small number (n=20) of zebrafish either injected with MO or a control (water) indicated that MO zebrafish may show a decreased heatrate. The average heartrate of MO injected zebrafish was 75 beats per minute (bpm), as opposed to 86bpm in controls (Figure 7.8).

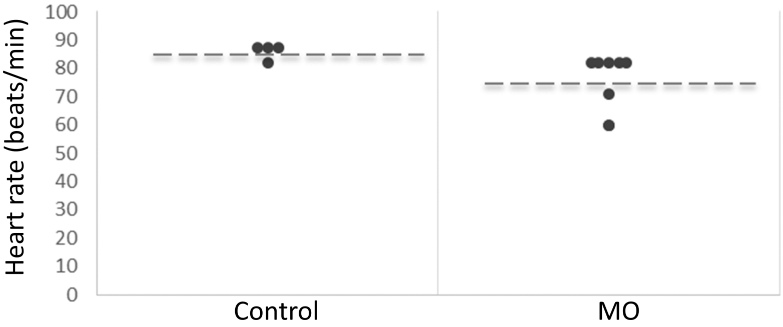


Figure 7.8 Zebrafish heartrates
Preliminary data on zebrafish embryo heartrates between MO treatment and untreated controls, indicating a decreased heartrate in MO fish. Average heartrate for each group is indicated by dashed line.

7.6 CRISPR/cas9 targeted genome editing

Based on the promising results from these preliminary morpholino data, it was decided that a stable *ppa2* knockout zebrafish line would be useful to confirm the morpholino phenotype, and further investigate the possible heart dysfunction and other relevant features. Additionally, *ppa2* sequences containing specific mutations of interest could then be introduced into the model, producing strains of fish that mimic the genotype seen in the human patients. In order to efficiently produce a *ppa2* knockout zebrafish line the CRISPR/cas9 system was utilised. The CRISPR/cas 9 system is a prokaryotic immune system that recognises and cuts exogenous genetic elements providing a form of acquired immunity. By synthesising a guide RNA that specifies a DNA target, and introducing this guide with cas9 the system can effectively be reprogrammed for targeted genome editing. Three guide RNA sequences were designed and introduced to zebrafish embryos with cas9 protein (100pm mRNA + 200pg cas9). At the time of writing these fish were being raised as fry.

Although it will take further breeding and genotyping to ascertain if a stable *ppa2* knockout line has been produced, many of these early fry will be mosaic ppa2 knock-outs and some early anecdotal observations are promising. At 2dpf, tail and fin defects are observed, mimicking the defects characterised in the morpholino experiments. At 4dpf, elongated 'beaky' jaws are observed, again mimicking a feature in the morpholino experiments (Figure 7.9). At the

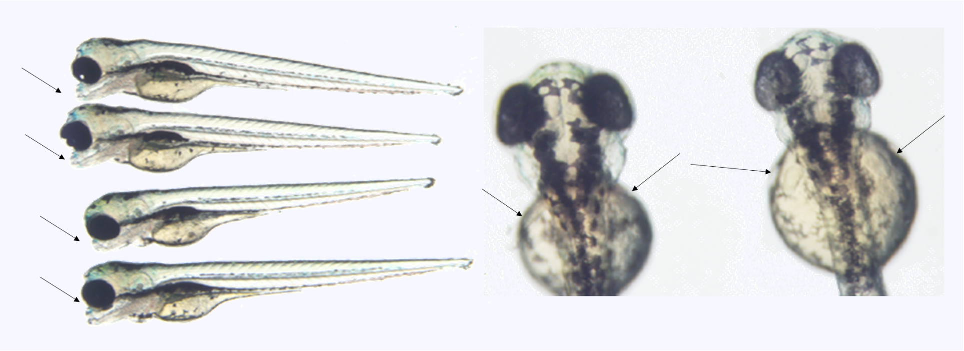


Figure 7.9 CRISPR/cas9 fry
Early mosaic fish phenotypes presenting with protruding jaws and absent or stunted pectoral fins.

time of writing, this work was still in preliminary stages, but further characterisation and genotyping of the next generation of zebrafish will be performed, alongside replicate experiments of morpholino injections in order to increase the data cohort to a statistically meaningful size.

7.7 Discussion

Although the experimentation in this chapter are ongoing and no final data is presented, this is largely only pending replication of experiments to increase zebrafish sample size.

Based on these preliminary results, it appears that injection of *ppa2*-targeted morpholinos results in stable phenotype that involves both cardiac and developmental features.

The severity of this phenotype is significantly reduced by co-injection of wildtype *ppa2* mRNA which acts to rescue normal *ppa2* expression in developing embryos. Two measures to quantify phenotype severity were used; a gross developmental score (based on set criteria), and a specific presence or absence of stunted pectoral fins. An expansion of these experiments aims to increase the sample size of these rescue experiments, and compare rescue success-rates when mutant *ppa2* mRNA is coinjected in place of wildtype mRNA.

Due to the potential for off target effects in morpholino knockdown experiments, the gold standard approach is still to produce a complete *ppa2* knockout model. The considerable time and cost involved in producing a knockout model made morpholino work a practical alternative to initially explore the effect of *ppa2* in zebrafish. There was also a concern that a *ppa2* knockout zebrafish model would be lethal in the homozygous state.

Given the promising preliminary results from the morpholino experiments, it was considered necessary to attempt to produce a knockout strain to validate the phenotype observed. CRISPR/cas9 was used for targeted genome edited in zebrafish embryos to produce mosaic *ppa2* knockout fry. The offspring of these fish will be genotyped to determine if germline incorporation occurred in some fish, and heterozyous *ppa2* knockout fish will become the breeding stock for a heterozygous *ppa2* knockout line.

The long term aim of producing a *ppa2* knockout line, is to eventually knock in wildtype and mutant human *PPA2* to observe the specific effects of mutations seen in human patients. In the short term, the knock out zebrafish will be used to repeat the phenotype characterisation and RNA rescue experiments previously performed in the morpholino fish. Wildtype and human *PPA2* mRNA will also be introduced to both morpholino and CRISPR/cas9 knockout fish in the RNA rescue experiments to ascertain if complementation will occur between the two gene homologs.

As with any experimental model, features seen in zebrafish with altered *ppa2* (or even *PPA2*) expression cannot be extrapolated to the human counterpart. Zebrafish *ppa2* may perform a subtly different function, and zebrafish are fundamentally different in development of structure such as gills, fins and tails. Bearing in mind these differences, it may still be possible to learn more about the function of *PPA2* in humans by observing the phenotype of morpholino KO and CRISPR/cas9 zebrafish models.

It should be noted that in the course of these studies, which were primarily directed at proving the phenotypic consequences of the compound heterozygous mutations within *PPA2*, other human families with *PPA2* mutations emerged in a group of European laboratories. The functional analysis of *PPA2* in yeast and zebrafish, although still very pertinent and biologically important, become less critical to the discovery of *PPA2* as causative of cardiac disease, and the functional studies of the variants were undertaken in patient fibroblasts and by enzymatic kinetic studies published by the European laboratories which we collaborated with and shared data. In actuality, Family Christchurch presented in the thesis, enabled the exome data stored in Europe, to be meaningful, as otherwise the variants may have remained VUS for a further period of time [1].

Chapter 8

Conclusions and future directions

8.1 Analysis of an intellectual disability pedigree

Despite exome analysis of three affected siblings in this pedigree, no genetic cause of their ID was identified. Due to the repetitive nature of the sex chromosomes, we were unable to exclude a causative feature on the Y-chromosome, and also encountered issues with mapping and alignment of reads on the sex chromosomes leading to some false variant calls. This case study proved invaluable in identifying some of the potential pitfalls associated with embarking on a new technology. Several recommendations (such as assessing X and Y chromosome coverage, and establishing a clear inheritance pattern where possible in a pedigree) for the implementation of exome sequencing in the clinical laboratory setting arose as a direct consequence of our experience during the analysis of this index case and subsequently their siblings. In this family, and future cases where an inheritance pattern is not certain, linkage association studies would be useful prior to launching into analysis exome data.

8.2 Trio analysis - a sporadic case of premature aging

A mutation was identified in BCOR which is putatively causative of some or all of the symptoms seen in this proband. This missense mutation is *de novo* in the proband, and occurs on the X chromosome. Two defined genetic disorders have been associated with mutations in BCOR, these are LM in males, and OFCD in females. The proband is not affected by OFCD, but shows symptoms that could be related to the function of BCOR. A common feature of pathogenic X-linked gene mutations in female carriers is skewed X-inactivation, preferentially expressing the X chromosome with a wildtype allele in tissues where this may confer a proliferation or survival benefit. Analysis of this individual showed moderately skewed inactivation (88:12) in peripheral blood, supporting the presence of a pathogenic X-linked mutation. Buccal and skin fibroblast DNA displayed normal X-inactivation.

An early avenue of investigation in this individual during childhood was to exlude a diagnosis of FA. Mitomycin c sensitivity of chromosomes from peripheral blood was measured, and returned a normal result. Given the recent identification of skewed X-inactivation in this patient, the mitomycin c assay for FA was repeated on cultured fibroblasts. This also returned a normal result, however there have been questions raised over possible confounding issues around the mitomycin c assay in the presence of BCOR mutations. A previous case of siblings with a familial nonsense mutation causing OFCD were also analysed for FA by mitomycin c sensitivity. Despite the same BCOR genotype and a defined phenotype, one sibling returned a positive result for FA. The phenotype was later identified as OFCD not FA but there may be potential for some overlap between these phenotypes in individuals with mutations in BCOR. As such, BCOR is considered a strong candidate for pathogenicity in this patient although the work needed to establish a causal link is beyond the scope of this PhD. More generally it would be interesting to follow up on the possible link between BCOR mutations and mitomycin C sensitivity.

8.3 Analysis of a pedigree with sudden cardiac death

Exome sequencing of affected siblings and their unaffected parents led to identification of compounding heterozygous mutations in PPA2. These mutations were predicted to be pathogenic by *in silico* algorithms and were in a mitochondrial-related gene, although PPA2 had not been previously described in human disease. This work led to a collaboration with several European laboratories, resulting in the identification of a further three families with compounding or homozygous mutations in PPA2 and a mitochondrial phenotype resulting in death during infancy [1]. The consistency of the phenotype in the three European families implies that mitochondrial symptoms and infantile death is likely to be the standard presentation of PPA2-associated disease. Functional experiments in yeast and zebrafish supported that normal PPA2 function is crucial to mitochondrial respiration, whereas decreased expression of PPA2 led to increased sensitivity to oxidative stress (yeast), developmental and growth issues of the heart and other 'high energy requirement' tissues (zebrafish). Continued work by collaborators supported this data, and provided evidence that the severity of p.Pro228Leu is less than the other PPA2 mutations identified. This explains the relatively mild phenotype seen in the original family, but raises concerns that other mild PPA2 mutations may put seemingly healthy individuals at risk of SCD. A mechanism of pathology for PPA2-associated disease has been hypothesised, based on what is known about the biochemistry of the mitochondria and the OXPHOS chain. Part of this mechanism is the exquisite alcohol sensitivity experienced in the original family. We hypothesise that alcohol is acting as a trigger of SCD, by significantly increasing the stress on mitochondria which are functinally sub optimallly.

Indentification of multiple unrelated families harbouring *PPA2* mutations, supported by functional data in yeast, zebrafish, recombinant *E coli* and patient fibroblasts have newly established this gene as pathogenic in the field of mitochondrial disorders. PPA2 dysfunction should now be considered in cases of neonatal or infantile death with symptoms of mitchondrial disease, and additionally may be relevant in SCD of apparently healthy individuals in the absence of more prevalent causes.

A mammalian model would provide another model which is a closer representation of human PPA2 function, and a Ppa2 knock-in mouse would be the logical next step. Several large scale commercial projects are currently underway to sequentially knockout mouse genes. During the course of this PhD, at least one company has made a Ppa2 constitutional knockout mouse model available for purchase (Taconic; http://www.taconic.com/). Further elucidation of the mechanism of PPA2 function, and in this case, alcohol metabolism and stress would be best provided with a mammalian model such as mice. Modelling would be crucial for increasing our understanding of this relatively unexplored enzyme and to give a system in which to test possible therapies.

8.4 Recommendations from thesis

Based on our experience with the MPS performed in this thesis, several learnings have emerged specific to our laboratory situation. Although exome sequencing has been utilised as a proof of concept in case studies presented here, exome sequencing will not be the focus of implementing MPS in our laboratory. Initially the priority would be to transfer and extend the Sanger sequencing currently performed to an MPS platform. A focused gene panel approach would be a more realistic introduction, to streamline the analysis burden in the first instance. Decisions on which genes, or groupings of disorders would be included on a panel will take into consideration the individual laboratory requirements such as turn-around time, read coverage, and sample volumes.

It may be beneficial to design multiple MPS panels which cover a group of genes associated with a general phenotype (such as hereditary deafness). In this scenario patient samples are only analysed for genes relevant to their condition. However, it may work out more economical and efficient to run all laboratory samples through a much larger panel incorporating genes associated with multiple phenotypes, then refine the analysis to only relevant genes. In this way, the laboratory will reduce the variants of uncertain significance that are identified, which currently don't contribute to a meaningful result for the clinician requesting testing. Work on defining which genes of interest will be included a panel, as well as looking at commercially available panels, is currently in process, and will like evolve beyond the course of this thesis. It is likely that exome or genome sequencing will be offered in the future, but this is dependant on a robust clinical pipeline for generating and analysing MPS which will be initially validated on MPS panels. More general learnings from the work produced in thesis are listed below.

1. Inheritance patterns (especially in striking families) are not always obvious. The assump-

tion that inheritance was X-linked or dominant in the cases of ID and SCD presented in this thesis was based on the high prevalence of affected individuals (5/5 males) and (4/4 offspring) respectively in these two families. Had we continued searching for dominant variants, and not considered all modes of inheritance in the family with SCD we would never have identified PPA2 as the causative gene in this disorder.

2. Many exome analyses will not yield a diagnosis, even when you have a clear phenotype and obvious inheritance. Exhaustive analysis of the case of ID presented in Chapter 3 did not reveal any promising candidate variants, and this experience proved the futility of continued analysis in the same dataset. ID cases have one of the highest proportion of no-diagnosis following exome sequencing, and this is only slightly increased by whole genome sequencing. While we are continuing to look for non-synonymous variants in coding regions of the genome, we will continue to miss diagnoses that have more subtle pathology.

3. Even missense mutations in functionally relevant genes (to the disorder of interest) can require extensive validation. Our work on PPA2 has taken 3.5 years and we have still not elucidated the mechanism of pathogenicity.

4. Our historical references of 'normal' sequences will always contain pathogenic variants, and therefore blanket exclusion of variants in a filtering process because they are included in dbSNP (for example) is going to result in false negative results. Although newer sources of population genetic data such as ExAC are more carefully curated, there will always be pathogenic variants documented.

5. You need to select the right approach for the diagnostic scenario. A disease targeted panel for cardiomyopathies would never have identified PPA2 variants in the family with SCD. In contrast, specific gene sequencing was entirely appropriate in two other clinical cases published during this thesis, which are not presented in the body of this work ([227, 228]).

6. It is important to have good clinical assessment data and phenotyping. Assessing relevance of candidate genes can be based entirely on the clinical description of the disorder involved, and if this is sparse you risk disregarding the relevant gene.

7. Always keep in mind the potential for mosaicism and sample mix up, results may not always be clear cut.

8.5 Summary

The process of working through the case studies of undiagnosed and complex genetic disease presented here has proved invaluable in gaining a understanding of how a technology such as massively parallel sequencing could be implemented in a clinical laboratory setting. The value of this analysis in cases of a defined phenotype with affected and unaffected family members is clear. Less certain is the approach we woudl take with singleton cases, as the variant filtering and analysis in this situation of dramatically less power, would require a much greater time investment than that of a highly powered family analysis with more than one affected individual. Although exome analysis has been used in all case studies in this thesis, the knowledge gained pertains to whole genome sequencing as well as targeted sequencing of gene panels. What was abundantly clear was that MPS was a very effective method of identifying disease candidates, not limited only to known causative genes. This presents the new dilemma of how far does the responsibility of the diagnostic lab stretch to prove pathogenicity of genetic variants detected which have not been previously catalogued. This thesis demonstrates how some characterisation of variants can be performed within the clinical setting, as well as via collaboration with research groups and suggests that characterisation, not identification, of genetic variants is the greatest challenge that diagnostic laboratories currently face.

8.6 Future directions

An unexpected outcome of this thesis was the discovery of an associated between *PPA2* and SCD, cardiomyopathy and other mitochondrial symptoms. As such, the key work expected to continue beyond the course of this PhD is to further characterise specific PPA2 mutations, elucidate a mechanism(s) of pathogenicity, and explain the heterogeneity seen between affected families, including the alcohol sensitivity seen in Family Christchurch.

Future areas of investigation are detailed here:

1. Create a stable heterozygous zebrafish CRISPR/cas9 knockout line (currently breeding, in conjunction with the Horsefield group, Otago Univesity).

2. Use CRISPR/cas9 technology to create a knock in of specific PPA2 variants in this line.

3. Continue mitochondrial function analysis of several cell types (monocytes, lymphocytes and fibroblasts) on the Seahorse XF instument (in conjunction with Andree Peason and Mark Hampton of the Free Radical group, University of Otago-Christchurch).

4.Use currently frozen fibroblasts and EBV transformed lymphocytes from affected and unaffected individuals (of Family Christchurch and age matched controls) to assess mitochondrial depletion, morphology of the mitochondria, levels of PPi in the mitochondrial matrix, mitochondrial activity when faced with ethanol stress, and other relevant experiments that may help to elucidate a mechanism of pathogenicity.

5. As mentioned in the above section, a mouse model is the final goal in order to be able to explore possible therapies for PPA2 dysfunction. Grant funding will be required for this project, and it will likely be a project taken on by a future student or post-doctoral researcher at the University of Otago.

Appendices

Appendix A

Primer Sequences

Table A.1 Primers used in general methods

Gene	Variant	Direction	Sequence	Comment
Universal		Forward	GTAAAACGACGGCCAGT	For sequencing M13 tagged PCR products
Universal		Reverse	CAGGAAACAGCTATGACC	For sequencing M13 tagged PCR products

Table A.2 Primers used in Chapter 3

Gene	Variant	Direction	Sequence	Comment
DACH2	p.Glu425Asp	Forward	GTAAAACGACGGCCAGTTGTGGCTTAGTGAGACTTTTGA	M13 Tagged
DACH2	p.Glu425Asp	Reverse	CAGGAAACAGCTATGACGGGAATCCAGCGGGCAATG	M13 Tagged
KMT2C	p.Gly838Ser	Forward	GTAAAACGACGGCCAGTAGACTTGCCTTCGCATGACA	M13 Tagged
KMT2C	p.Gly838Ser	Reverse	CAGGAAACAGCTATGACCCTACCTGTTTGGACCGAGG	M13 Tagged
KMT2C	p.Leu291Phe	Forward	GTAAAACGACGGCCAGTTACCCCTTTCAGCGATGTGC	M13 Tagged
KMT2C	p.Leu291Phe	Reverse	CAGGAAACAGCTATGACCAATGTGTTCTGGACAAAGC	M13 Tagged
KMT2C	p.Gly838Ser	Forward	GTAAAACGACGGCCAGTCTTCAGACTTGCCTTCGCATGACATGCTGGA	Specific for KMT2C
KMT2C	p.Gly838Ser	Reverse	CAGGAAACAGCTATGACCCAACTGAGTTGAAATTCCAAATAAGATGA	Specific for KMT2C
KMT2C	p.Leu291Phe	Forward	GTAAAACGACGGCCAGTTTGCAGGTTCAATTTGTTTTACTTGTGATA	Specific for KMT2C
KMT2C	p.Leu291Phe	Reverse	CAGGAAACAGCTATGACAAATACATTGGTTGAAGTTTAAAAGTCTATC	Specific for KMT2C

Table A.3 Primers used in Chapter 4

Gene	Variant	Direction	Sequence	Comment
BCOR	p.Gly1303Ser	Forward	GTAAAACGACGGCCAGTCTGGCTACAACTGTGTAACG	
BCOR	p.Gly1303Ser	Reverse	CAGGAAACAGCTATGACCACAGCACATCTGTCTTCTGG	
BCOR	p.Gly1303Ser	Forward	TCCCTGCACTGGTGGATGAAAGACTCTTC	Mutation specific nest primer
BCOR	p.Gly1303Gly	Forward	CCCCTGCACTGGTGGATGAAAGACTCTTC	WT specific nest primer
AR	n/a	Forward	/56-FAM/TCCAGAATCTGTTCCAGAGCGTGC	
AR	n/a	Reverse	GCTGTGAAGGTTGCTGTTCCTCAT	

Table A.4 Primers used in Chapters 5-6

Gene	Variant	Direction	Sequence	Comment
PPA2	p.Glu172Leu	Forward	TGGCATGAATAAACCAAAACTCACG	PCR/Seq
PPA2	p.Glu172Leu	Reverse	GACTTGGGAAGATCCCCATGAAA	PCR/Seq
PPA2	p.Pro228Leu	Forward	AGATTCTCATCTTCACCTTG	PCR/Seq
PPA2	p.Pro228Leu	Reverse	TTGTGCTTCTGGTTAGACTG	PCR/Seq

Appendix B

Zebrafish experimental details

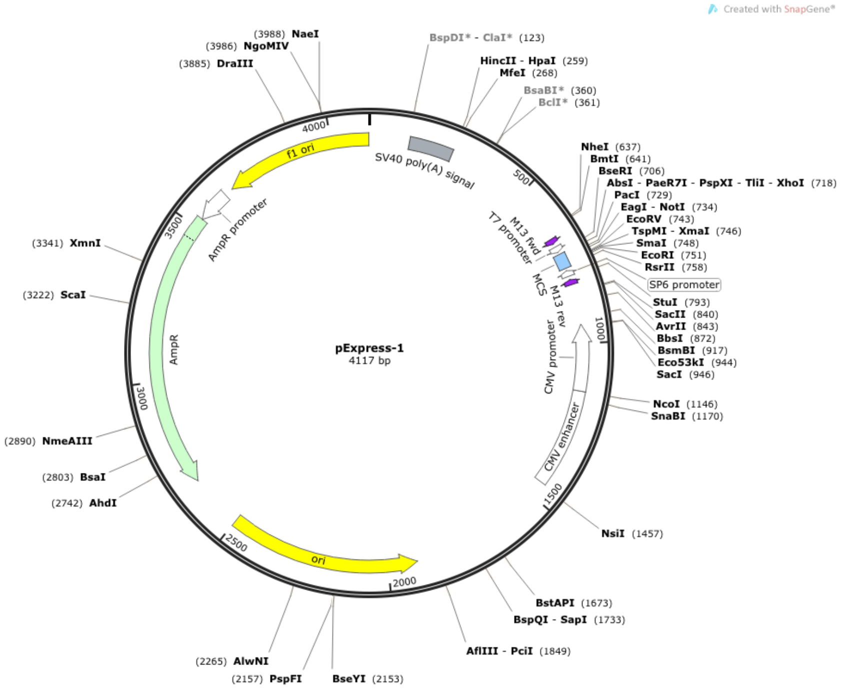


Figure B.1 pExpress-1
Vector map of pExpress-1 showing cloning sites.

Table B.1 Morpholino sequences
Used for transient knockdown of ppa2 expression.

Construct	Construct Name	MO Sequence
1	PPA2 5'UTR	CTGAACACATTATTTACATCCACTC
2	PPA2 SPX	GCAAAAACAGAACATCACCTGTGGA
3	PPA2 MIS-5	GAAAAAAAAGAAAATCACCTATGAA

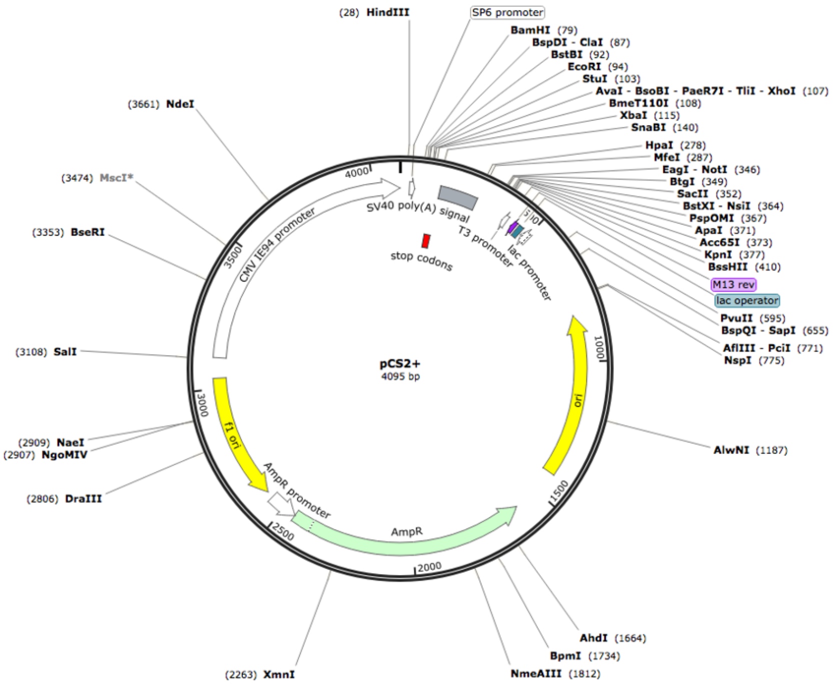


Figure B.2 pCS2+ Plasmid
Vector map of pCS2 showing cloning sites.

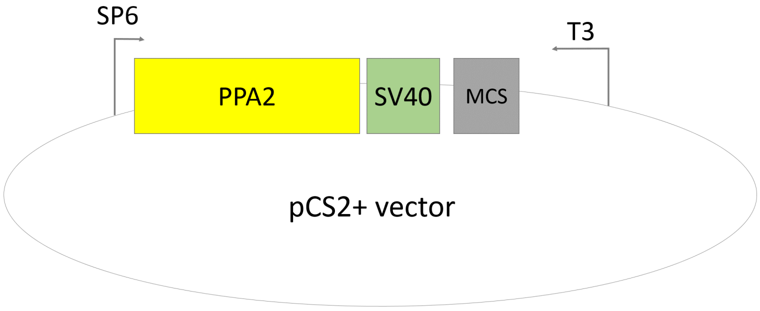


Figure B.3 pCS2+ cloning plan

Appendix C

Gene lists used in exome analysis

Table C.1 Intellectual disability (ID) associated genes

Gene list examined against variants identified during exome analysis of the pedigree in Chapter 3.

ABCC6
ABCD1
ABCG5
ACAT1
ACOX1
ACSL4
ACY1
ADAR
ADSL
AFF2
AFP
AGL
AGT
AGTR2
AHI1
AIFM1
ALDH18A1
ALDH4A1
ALG11
ALG12
ALG6
ALX4
AMER1
ANK3
AP1S1
AP1S2
AP3B1
AP4B1
AP4E1
AP4M1
AP4S1
APOB
AQP7
AR
ARFGEF2
ARG1
ARHGEF6
ARHGEF9
ARID1A
ARID1B
ARL6
ARX
ASPM
ASS1
ATL1
ATM
ATP13A2
ATP1A2
ATP6AP2
ATP7A
ATRX
AUH
AVP
AVPR1A
AVPR2
BBIP1
BBS2
BBS4
BBS5
BBS7
BBS9
BBS10
BBS12
BBS14
BBS18
BCOR
BCS1L
BDNF
BIN1
BRAF
BRCA2
BRIP1
BRWD3
BUB1B
C12orf57
CA8
CACNA1C
CACNG2
CAMTA1
CANT1
CASK
CBL
CBS
CC2D1A
CC2D2A
CCDC22
CCDC88C
CDH7
CDH8
CDH15
CDKL5
CDKN1C
CEP290
CEP41
CEP57
CHD7
CDH15
CDKL5
CHRNA4
CLIC2
CLN3
CNTNAP2
COG5
COG7
COL1A2
CP
CPA6
CPS1
CRADD
CREBBP
CRBN
CTC1
CTCF
CTNNB1
CTSA
CUL4B
CYB5R3
CYP27A1
D2HGDH
DARS2
DBT
DCX
DDHD2
DHCR24
DHCR7
DIP2B
DLG3
DMD
DPYD
DYNC1H1
DYRK1A
EBP
EFNB1
EHMT1
EIF2S3
ELOVL4
EPB41L1
ERCC2
ERCC3
ERCC5
ERCC6
ERCC8
ERLIN2
EZH2
F5
FAM126A
FANCG
FBLN5
FBN1
FBN2
FGD1
FGF14
FGFR1
FGFR2
FGFR3
FKRP
FKTN
FLNA
FMR1
FOLR1
FOXG1
FOXP1
FOXP2
FRMPD4
FTO
FTSJ1
G6PC3
GABRG2
GALE
GAMT
GAN
GATAD2B
GBA
GBE1
GCK
GDI1
GFAP
GFM1
GHR
GLI3
GLRA1
GLUL
GLYCTK
GM2A
GNAS
GNPAT
GNPTAB
GNPTG
GPC3
GRIA3
GRIK2
GRIN1
GRIN2A
GRIN2B
GRM1
GSS
GUSB
GYS2
HAX1
HCCS
HCFC1
HDAC4
HDAC8
HEPACAM
HEXB
HOXA1
HOXD10
HPD
HPRT1
HRAS
HSD17B10
HSPD1
HUWE1
IDH2
IDS
IGBP1
IGF1
IGF1R
IL1RAPL1
INSR
IQSEC2
IRX5
ITGA7
KCNJ10
KCNJ11
KCNK9
KCNQ2
KCTD7
KDM5C
KIAA2022
KMT2D
KIF11
KIF1A
KIF21A
KIF5A
KIF7
KIRREL3
KLF8
KRAS
L1CAM
L2HGDH
LAMA2
LAMP2
LARGE
LBR
LHX3
LIG4
LINS
LMBRD1
LRP5
LYST
LZTFL1
MAGT1
MAN1B1
MAN2B1
MANBA
MAOA
MAP2K1
MAP2K2
MAPT
MAT1A
MBD5
MBTPS2
MCCC1
MCCC2
MCOLN1
MCPH1
MECP2
MED12
MED17
MED23
MEF2C
MET
MFSD8
MGAT2
MID1
MKKS
MKS1
MLL2
MMADHC
MOCS2
MPI
MPZ
MRAP
MTFMT
MTHFR
MTR
MYCN
MYO5A
MYO7A
NAA10
NAGA
NBN
NDP
NDUFA1
NDUFAF5
NDUFS1
NF1
NFIX
NGF
NHEJ1
NHP2
NHS
NIPBL
NLGN3
NLGN4X
NPC1
NPC2
NPHP3
NRAS
NRXN1
NSD1
NSDHL
NSUN2
OCRL
OFD1
OPHN1
ORC1
PACS1
PAFAH1B1
PAH
PAK3
PAX6
PCDH19
PCNT
PDE4D
PDHA1
PDHX
PDSS1
PEX7
PGK1
PHF6
PHF8
PHKA2
PHKG2
PIGL
PIGO
PIGV
PLA2G6
PLP1
PNKP
POMGNT1
POMT1
POMT2
POU1F1
PPOX
PQBP1
PRICKLE1
PRKAR1A
PRPS1
PRSS12
PTCHD1
PTEN
PTPN11
PYCR1
PYGL
RAB39B
RAB40AL
RAD21
RAF1
RAI1
RAPSN
RBBP8
RBM10
RELM
RFX6
RPGRIP1L
RPL10
RPS6KA3
SACS
SAMHD1
SATB2
SCN1A
SCN2A
SCN8A
SDCCAG8
SGCA
SGSH
SHANK2
SHANK3
SHOC2
SHROOM4
SIL1
SLC16A2
SLC20A2
SLC25A1
SLC25A12
SLC25A13
SLC25A15
SLC2A1
SLC2A2
SLC35C1
SLC46A1
SLC4A4
SLC5A2
SLC5A5
SLC6A4
SLC6A8
SLC7A7
SLC9A6
SLX4
SMARCA4
SMARCB1
SMARCE1
SMC1A
SMC3
SMS
SNIP1
SOBP
SOS1
SOX10
SOX2
SOX3
SPR
SPRED1
SPTAN1
SPTLC1
SRD5A3
SRPX2
ST3GAL3
STAT5B
STRA6
STX11
STXBP1
SUCLG1
SYN1
SYNGAP1
SYP
SYT14
TBC1D24
TBCE
TBX1
TCF4
TECR
TGIF1
TH
THRB
TINF2
TMCO1
TMEM165
TMEM216
TMEM67
TMEM70
TPH2
TPK1
TRAPPC9
TRHR
TRIM32
TSC1
TSC2
TSHR
TSPAN7
TTC8
TTC37
TTR
TUBA1A
TUBA8
TUBB2B
TUBB3
TUSC3
TWIST1
UBE2A
UBE3A
UPB1
UPF3B
UROC1
USP9X
VLDLR
VPS13B
WDPCP
WDR62
WDR81
WRN
XIST
XPNPEP3
ZBTB16
ZBTB24
ZC3H14
ZDHHC9
ZDHHC15
ZEB2
ZFP57
ZFYVE26
ZIC2
ZNF41
ZNF81
ZNF526
ZNF674
ZNF711

Table C.2 Autism spectrum disorder (ASD) associated genes

Gene list examined against variants identified during exome analysis of the pedigree in Chapter 3.

ABAT
ABCA7
ACTN4
ADA
ADAMTS18
ADARB1
ADCY5
ADK
ADNP
ADORA2A
ADORA3
ADRB2
AFF4
AGAP1
AGBL4
AGMO
ALDH1A3
ALDH5A1
ALOX5AP
AMPD1
AMT
ANK2
ANKRD11
ANKS1B
ANXA1
APBA2
APC
APH1A
APP
ARHGAP11B
ARHGAP15
ARHGAP24
ARNT2
ASH1L
ASMT
ASTN2
ASXL3
ATG7
ATP10A
ATP2B2
ATRNL1
ATXN7
AUTS2
BAIAP2
BCAS1
BCKDK
BCL11A
BCL2
BIRC6
BTAF1
BZRAP1
C11orf30
C15orf43
C3orf58
C4B
CA6
CACNA1B
CACNA1D
CACNA1F
CACNA1G
CACNA1H
CACNA1I
CACNA2D3
CACNB2
CADM1
CADM2
CADPS2
CAMK4
CAMSAP2
CAPRIN1
CASC4
CCDC64
CD38
CD44
CD99L2
CDC42BPB
CDH10
CDH11
CDH22
CDH9
CDKN1B
CECR2
CELF4
CELF6
CHD2
CHD7
CHRM3
CHRNA7
CHRNB3
CHST5
CIB2
CLSTN3
CLTCL1
CMIP
CNR1
CNR2
CNTN3
CNTN4
CNTN5
CNTN6
CNTNAP3
CNTNAP5
CSMD1
CSNK1D
CTNNA3
CTTNBP2
CUL3
CX3CR1
CXCR3
CYFIP1
CYP11B1
DAB1
DAGLA
DAPK1
DAPP1
DCTN5
DCUN1D1
DDC
DDX11
DDX53
DEAF1
DEPDC5
DGKZ
DIAPH3
DIP2A
DISC1
DLG1
DLG4
DLGAP1
DLGAP2
DLGAP3
DLX1
DLX2
DLX6
DMPK
DNAH10
DNAJC19
DNER
DNM1L
DOCK10
DOCK4
DOLK
DPP6
DRD1
DRD2
DRD3
DSCAM
DST
DUSP22
DYDC1
DYDC2
EEF1A2
EFR3A
EGR2
EIF4E
EIF4EBP2
EML1
EN2
EP300
EP400
EPC2
EPHA6
EPHB2
EPHB6
EPS8
ERBB2IP
ERBB4
ERG
ESR1
ESR2
ESRRB
ETFB
EXOC3
EXOC5
EXOC6
EXOC6B
EXT1
F13A1
FABP3
FABP5
FABP7
FAM135B
FAN1
FAT1
FBXO15
FBXO33
FBXO40
FER
FEZF2
FGA
FGFBP3
FHIT
FLT1
FOLH1
FRK
GABRA1
GABRA3
GABRA4
GABRB1
GABRB3
GABRQ
GAD1
GALNT13
GALNT14
GAP43
GAS2
GATM
GDA
GIGYF1
GLO1
GLRA2
GNA14
GNB1L
GPC6
GPD2
GPHN
GPR139
GPR37
GPX1
GRID1
GRID2
GRID2IP
GRIK3
GRIP1
GRM4
GRM5
GRM8
GRPR
GSK3B
GSN
GSTM1
GTF2I
GUCY1A2
HCN1
HDAC6
HERC2
HLA-A
HLA-DRB1
HMGN1
HNRNPH2
HOMER1
HOXB1
HS3ST5
HSD11B1
HTR1B
HTR2A
HTR3A
HTR3C
HTR7
HYDIN
ICA1
IL16
IL1R2
IL1RAPL2
IMMP2L
INADL
INPP1
ITGA4
ITGB3
ITGB7
JARID2
JMJD1C
KANK1
KATNAL2
KCND2
KCNJ12
KCNJ15
KCNJ2
KCNMA1
KCNQ3
KCNT1
KCTD13
KDM5B
KDM6B
KHDRBS2
KIAA1586
KIF13B
KIF5C
KIT
KLC2
KMO
KMT2A
KMT2C
KMT2E
KPTN
LAMA1
LAMB1
LAMC3
LEP
LIN7B
LMX1B
LPL
LRFN5
LRP2
LRP2BP
LRPPRC
LRRC1
LRRC7
LZTS2
MACROD2
MAGED1
MAGEL2
MAL
MAOB
MAP2
MAPK1
MAPK12
MAPK3
MAPK8IP2
MARK1
MBD1
MBD3
MBD4
MBD6
MC4R
MCC
MDGA2
MED13L
MIB1
MKL2
MPP6
MSN
MSNP1AS
MSR1
MTF1
MTX2
MYH4
MYO1A
MYO9B
MYOZ1
MYT1L
NAA15
NBEA
NCKAP1
NCKAP5
NCKAP5L
NDNL2
NDUFA5
NEFL
NELL1
NFIA
NIPA1
NIPA2
NLGN1
NLGN2
NLGN4Y
NOS1
NOS1AP
NOS2A
NPAS2
NR3C2
NRCAM
NRG1
NRP2
NRXN2
NRXN3
NTNG1
NTRK1
NTRK3
NXF5
NXPH1
ODF3L2
OGT
OPRM1
OR1C1
OTX1
OXTR
P2RX4
PARD3B
PARK2
PAX5
PCDH10
PCDH15
PCDH8
PCDH9
PCDHA1
PCDHA10
PCDHA11
PCDHA12
PCDHA13
PCDHA2
PCDHA3
PCDHA4
PCDHA5
PCDHA6
PCDHA7
PCDHA8
PCDHA9
PCDHAC1
PCDHAC2
PCDHGA11
PDE1C
PDE4A
PDE4B
PDZD4
PECR
PER1
PHF2
PIK3CG
PIK3R2
PINX1
PITX1
PLAUR
PLCB1
PLCD1
PLN
PLXNA4
POGZ
PON1
POT1
PPFIA1
PPP1R1B
PPP1R3F
PRICKLE2
PRKCB
PRKD1
PRODH
PRSS38
PRUNE2
PSD3
PSMD10
PTGER3
PTGS2
PTPRB
PTPRC
PTPRT
PXDN
RAB11FIP5
RAB19
RAPGEF4
RASD1
RASSF5
RB1CC1
RBFOX1
RBM8A
RBMS3
REEP3
RELN
RERE
RFWD2
RGS7
RHOXF1
RIMS1
RIMS3
RNPS1
ROBO1
ROBO2
RORA
RPP25
RPS6KA2
SAE1
SBF1
SCFD2
SCN5A
SCN7A
SDC2
SDK1
SEMA5A
SERPINE1
SETBP1
SETD2
SETD5
SETDB1
SETDB2
SEZ6L2
SGSM3
SH3KBP1
SHANK1
SLC16A3
SLC16A7
SLC1A1
SLC1A2
SLC22A15
SLC24A2
SLC25A14
SLC25A24
SLC25A27
SLC29A4
SLC30A5
SLC35A3
SLC38A10
SLC39A11
SLC4A10
SLC6A3
SLC9A9
SLCO1B3
SLIT3
SLITRK5
SMG6
SNAP25
SND1
SNRPN
SNTG2
SNX19
SOD1
SOX5
SPAST
SRD5A2
SRRM4
ST7
ST8SIA2
STK39
STX1A
STXBP5
SUCLG2
SUV420H1
SYAP1
SYN2
SYN3
SYNE1
SYT17
SYT3
TAF1C
TAF1L
TBC1D5
TBC1D7
TBL1X
TBL1XR1
TBR1
TCF20
TCF7L2
TDO2
TGM3
THAP8
THBS1
THRA
TLK2
TM4SF20
TMEM231
TMLHE
TNIP2
TNN
TNRC6B
TOMM20
TOP1
TOP3B
TPO
TRIM33
TRIO
TRIP12
TRPC6
TRPM1
TSN
TTI2
TTN
TUBGCP5
TYR
UBA6
UBE2H
UBE3C
UBL7
UBR5
UBR7
UPF2
UPF2
USP9Y
UTRN
VASH1
VIL1
VIP
WAC
WDFY3
WDR93
WNK3
WNT1
WNT2
XIRP1
XPC
XPO1
YEATS2
YTHDC2
YWHAE
ZBTB20
ZMYND11
ZNF18
ZNF385B
ZNF407
ZNF517
ZNF713
ZNF774
ZNF8
ZNF804A
ZNF827
ZSWIM5

Bibliography

[1] H. Kennedy, T. B. Haack, V. Hartill, L. Mataković, E. R. Baumgartner, H. Potter, R. Mackay, C. L. Alston, S. OSullivan, R. McFarland, *et al.*, "Sudden cardiac death due to deficiency of the mitochondrial inorganic pyrophosphatase PPA2," *The American Journal of Human Genetics*, vol. 99, no. 3, pp. 674–682, 2016.

[2] W. H. Organization *et al.*, "Genomics and world health: Report of the Advisory Committee on Health Research," 2002.

[3] S. E. McCandless, J. W. Brunger, and S. B. Cassidy, "The burden of genetic disease on inpatient care in a childrens hospital," *The American Journal of Human Genetics*, vol. 74, no. 1, pp. 121–127, 2004.

[4] D. N. Cooper, J.-M. Chen, E. V. Ball, K. Howells, M. Mort, A. D. Phillips, N. Chuzhanova, M. Krawczak, H. Kehrer-Sawatzki, and P. D. Stenson, "Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics," *Human Mutation*, vol. 31, no. 6, pp. 631–655, 2010.

[5] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. MacKenzie, "Rare-disease genetics in the era of next-generation sequencing: discovery to translation," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 681–691, 2013.

[6] P. S. Frenette, G. F. Atweh, *et al.*, "Sickle cell disease: old discoveries, new concepts, and future promise," *The Journal of Clinical Investigation*, vol. 117, no. 4, pp. 850–858, 2007.

[7] J. Hunt and V. Ingram, "Allelomorphism and the chemical differences of the human haemoglobins A, S and C," 1958.

[8] S. A. Sarnaik, "Thalassemia and related hemoglobinopathies," *The Indian Journal of Pediatrics*, vol. 72, no. 4, pp. 319–324, 2005.

[9] R. Bernards and R. A. Flavell, "Physical mapping of the globin gene deletion in hereditary persistence of foetal haemoglobin (HPFH)," *Nucleic Acids Research*, vol. 8, no. 7, pp. 1521–1534, 1980.

[10] R. J. Roberts, "How restriction enzymes became the workhorses of molecular biology," *Proceedings of the National Academy of Sciences*, vol. 102, no. 17, pp. 5905–5908, 2005.

[11] C. Camaschella and G. Saglio, "Recent advances in diagnosis of hemoglobinopathies," *Critical reviews in Oncology/Hematology*, vol. 14, no. 2, pp. 89–105, 1993.

[12] D. Nathans and H. O. Smith, "Restriction endonucleases in the analysis and restructuring of DNA molecules," *Annual review of Biochemistry*, vol. 44, no. 1, pp. 273–293, 1975.

[13] D. A. Jackson, R. H. Symons, and P. Berg, "Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 69, no. 10, pp. 2904–2909, 1972.

[14] J. E. Mertz and R. W. Davis, "Cleavage of DNA by R1 restriction endonuclease generates cohesive ends," *Proceedings of the National Academy of Sciences*, vol. 69, no. 11, pp. 3370–3374, 1972.

[15] P. E. Lobban and A. Kaiser, "Enzymatic end-to-end joining of DNA molecules," *Journal of Molecular Biology*, vol. 78, no. 3, pp. 453–471, 1973.

[16] S. N. Cohen, A. C. Chang, H. W. Boyer, and R. B. Helling, "Construction of biologically functional bacterial plasmids in vitro," *Proceedings of the National Academy of Sciences*, vol. 70, no. 11, pp. 3240–3244, 1973.

[17] Y. W. Kan, M. S. Golbus, and A. M. Dozy, "Prenatal diagnosis of α-thalassemia: clinical application of molecular hybridization," *New England Journal of Medicine*, vol. 295, no. 21, pp. 1165–1167, 1976.

[18] M. Farrall, C. Rodeck, P. Stanier, W. Lissens, E. Watson, H.-Y. Law, R. Warren, M. Super, P. Scambler, B. Wainwright, *et al.*, "First-trimester prenatal diagnosis of cystic fibrosis with linked DNA probes," *The Lancet*, vol. 327, no. 8495, pp. 1402–1405, 1986.

[19] S. L. Woo, A. S. Lidsky, F. Güttler, T. Chandra, and K. J. Robson, "Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria," *Nature*, vol. 306, no. 5939, pp. 151–155, 1983.

[20] I. S. Johnson, "Human insulin from recombinant DNA technology," *Science*, vol. 219, no. 4585, pp. 632–637, 1983.

[21] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim, "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia," *Science*, vol. 230, no. 4732, pp. 1350–1354, 1985.

[22] K. B. Mullis and F. A. Faloona, "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction.," *Methods in Enzymology*, vol. 155, p. 335, 1987.

[23] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich, "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase," *Science*, vol. 239, no. 4839, pp. 487–491, 1988.

[24] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences*, vol. 74, no. 2, pp. 560–564, 1977.

[25] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[26] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis," *Nature*, vol. 321, pp. 674–679, 1986.

[27] B. Rosenblum, L. Lee, S. Spurgeon, S. Khan, S. Menchen, C. Heiner, and S. Chen, "New dye-labeled terminators for improved DNA sequencing patterns," *Nucleic acids research*, vol. 25, no. 22, pp. 4500–4504, 1997.

[28] A. I. Tauber and S. Sarkar, "The ideology of the human genome project.," *Journal of the Royal Society of Medicine*, vol. 86, no. 9, p. 537, 1993.

[29] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[30] J. C. Venter, 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, *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[31] I. H. G. S. Consortium *et al.*, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.

[32] A. Goffeau, B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, *et al.*, "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[33] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, *et al.*, "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[34] S. E. Celniker, D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S. P. Dugan, E. Frise, *et al.*, "Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence," *Genome Biology*, vol. 3, no. 12, p. research0079, 2002.

[35] S. Consortium *et al.*, "Genome sequence of the nematode C. elegans: A platform for investigating biology," *Science*, vol. 282, pp. 2012–2018, 1998.

[36] A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, *et al.*, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–562, 2002.

[37] R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, *et al.*, "Genome sequence of the brown norway rat yields insights into mammalian evolution," *Nature*, vol. 428, no. 6982, pp. 493–521, 2004.

[38] K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, M. C. Zody, *et al.*, "Genome sequence, comparative analysis and haplotype structure of the domestic dog," *Nature*, vol. 438, no. 7069, pp. 803–819, 2005.

[39] L. Hillier, W. Miller, E. Birney, W. Warren, R. Hardison, C. Ponting, P. Bork, D. Burt, M. Groenen, M. Delany, *et al.*, "International chicken genome sequencing consortium: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution," *Nature*, vol. 432, pp. 695–716, 2004.

[40] T. C. Sequencing, A. Consortium, *et al.*, "Initial sequence of the chimpanzee genome and comparison with the human genome," *Nature*, vol. 437, no. 7055, pp. 69–87, 2005.

[41] R. A. Gibbs, J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, R. K. Wilson, *et al.*, "Evolutionary and biomedical insights from the rhesus macaque genome," *Science*, vol. 316, no. 5822, pp. 222–234, 2007.

[42] C. G. Elsik, R. L. Tellam, K. C. Worley, *et al.*, "The genome sequence of taurine cattle: a window to ruminant biology and evolution," *Science*, vol. 324, no. 5926, pp. 522–528, 2009.

[43] E. R. Mardis, "A decade/'s perspective on DNA sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, 2011.

[44] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Research*, p. gkp552, 2009.

[45] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, *et al.*, "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[46] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[47] I. C. for Blood Pressure Genome-Wide Association Studies *et al.*, "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[48] B. A. Bejjani and L. G. Shaffer, "Application of array-based comparative genomic hybridization to clinical diagnostics," *The Journal of Molecular Diagnostics*, vol. 8, no. 5, pp. 528–533, 2006.

[49] M. Doornbos, B. Sikkema-Raddatz, C. A. Ruijvenkamp, T. Dijkhuizen, E. K. Bijlsma, A. C. Gijsbers, Y. Hilhorst-Hofstee, R. Hordijk, K. T. Verbruggen, W. M. Kerstjens-Frederikse, *et al.*, "Nine patients with a microdeletion 15q11. 2 between breakpoints 1 and 2 of the Prader–Willi critical region, possibly associated with behavioural disturbances," *European Journal of Medical Genetics*, vol. 52, no. 2, pp. 108–115, 2009.

[50] D. A. Koolen, W. Reardon, E. M. Rosser, D. Lacombe, J. A. Hurst, C. J. Law, E. M. Bongers, C. M. van Ravenswaaij-Arts, M. A. Leisink, A. G. van Kessel, *et al.*, "Molecular characterisation of patients with subtelomeric 22q abnormalities using chromosome specific array-based comparative genomic hybridisation," *European Journal of Human Genetics*, vol. 13, no. 9, pp. 1019–1024, 2005.

[51] L. G. Shaffer, C. D. Kashork, R. Saleki, E. Rorem, K. Sundin, B. C. Ballif, and B. A. Bejjani, "Targeted genomic microarray analysis for identification of chromosome abnormalities in 1500 consecutive clinical cases," *The Journal of Pediatrics*, vol. 149, no. 1, pp. 98–102, 2006.

[52] M. J. Somerville, C. B. Mervis, E. J. Young, E.-J. Seo, M. del Campo, S. Bamforth, E. Peregrine, W. Loo, M. Lilley, L. A. Pérez-Jurado, *et al.*, "Severe expressive-language delay related to duplication of the Williams–Beuren locus," *New England Journal of Medicine*, vol. 353, no. 16, pp. 1694–1701, 2005.

[53] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, *et al.*, "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, no. 7189, pp. 872–876, 2008.

[54] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi, "Next-generation sequencing: from basic research to diagnostics," *Clinical Chemistry*, vol. 55, no. 4, pp. 641–658, 2009.

[55] D. R. Bentley, 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, *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 7218, pp. 53–59, 2008.

[56] E. R. Mardis, "Next-generation sequencing platforms," *Annual review of Analytical Chemistry*, vol. 6, pp. 287–303, 2013.

[57] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, *et al.*, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[58] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer, "Characterization of individual polynucleotide molecules using a membrane channel," *Proceedings of the National Academy of Sciences*, vol. 93, no. 24, pp. 13770–13773, 1996.

[59] I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao, M. Pavlenok, M. Niederweis, and J. H. Gundlach, "Nanopore DNA sequencing with MspA," *Proceedings of the National Academy of Sciences*, vol. 107, no. 37, pp. 16060–16065, 2010.

[60] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nature Nanotechnology*, vol. 4, no. 4, pp. 265–270, 2009.

[61] S. McGinn and I. G. Gut, "DNA sequencing–spanning the generations," *New Biotechnology*, vol. 30, no. 4, pp. 366–372, 2013.

[62] H.-C. Wu, Y. Astier, G. Maglia, E. Mikhailova, and H. Bayley, "Protein nanopores with covalently attached molecular adapters," *Journal of the American Chemical Society*, vol. 129, no. 51, pp. 16142–16148, 2007.

[63] E. A. Manrao, I. M. Derrington, A. H. Laszlo, K. W. Langford, M. K. Hopper, N. Gillgren, M. Pavlenok, M. Niederweis, and J. H. Gundlach, "Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase," *Nature Biotechnology*, vol. 30, no. 4, pp. 349–353, 2012.

[64] G. M. Cherf, K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus, and M. Akeson, "Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision," *Nature Biotechnology*, vol. 30, no. 4, pp. 344–348, 2012.

[65] A. S. Mikheyev and M. M. Tin, “A first look at the Oxford Nanopore minION sequencer,” *Molecular Ecology Resources*, vol. 14, no. 6, pp. 1097–1102, 2014.

[66] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, “Improved data analysis for the MinION nanopore sequencer,” *Nature Methods*, 2015.

[67] J. M. Urban, J. Bliss, C. E. Lawrence, and S. A. Gerbi, “Sequencing ultra-long DNA molecules with the Oxford Nanopore minION,” *BioRxiv*, p. 019281, 2015.

[68] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, *et al.*, “The potential and challenges of nanopore sequencing,” *Nature Biotechnology*, vol. 26, no. 10, pp. 1146–1153, 2008.

[69] C. Larsson, I. Grundberg, O. Söderberg, and M. Nilsson, “In situ detection and genotyping of individual mRNA molecules,” *Nature Methods*, vol. 7, no. 5, pp. 395–397, 2010.

[70] S. H. Katsanis and N. Katsanis, “Molecular genetic testing and the future of clinical genomics,” *Nature Reviews Genetics*, vol. 14, no. 6, pp. 415–426, 2013.

[71] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner, “Target-enrichment strategies for next-generation sequencing,” *Nature Methods*, vol. 7, no. 2, pp. 111–118, 2010.

[72] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, *et al.*, “Targeted capture and massively parallel sequencing of 12 human exomes,” *Nature*, vol. 461, no. 7261, pp. 272–276, 2009.

[73] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, *et al.*, “Exome sequencing identifies the cause of a mendelian disorder,” *Nature Genetics*, vol. 42, no. 1, pp. 30–35, 2010.

[74] A. Hoischen, B. W. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, *et al.*, “De novo mutations of SETBP1 cause Schinzel-Giedion syndrome,” *Nature Genetics*, vol. 42, no. 6, pp. 483–485, 2010.

[75] S. B. Ng, D. A. Nickerson, M. J. Bamshad, and J. Shendure, “Massively parallel sequencing and rare disease,” *Human Molecular Genetics*, vol. 19, no. R2, pp. R119–R124, 2010.

[76] J. H. Schuurs-Hoeijmakers, M. T. Geraghty, E.-J. Kamsteeg, S. Ben-Salem, S. T. de Bot, B. Nijhof, I. I. van de Vondervoort, M. van der Graaf, A. C. Nobau, I. Otte-Höller, *et al.*, “Mutations in DDHD2, encoding an intracellular phospholipase A, cause a recessive form of complex Hereditary Spastic Paraplegia,” *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 1073–1081, 2012.

[77] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski, “A survey of tools for variant analysis of next-generation genome sequencing data,” *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 256–278, 2014.

[78] M. Dai, R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan, D. M. Markovitz, G. Omenn, and F. Meng, “NGSQC: cross-platform quality analysis pipeline for deep sequencing data,” *BMC Genomics*, vol. 11, no. Suppl 4, p. S7, 2010.

[79] R. Schmieder and R. Edwards, “Quality control and preprocessing of metagenomic datasets,” *Bioinformatics*, vol. 27, no. 6, pp. 863–864, 2011.

[80] D. Blankenberg, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko, *et al.*, “Manipulation of FASTQ data with Galaxy,” *Bioinformatics*, vol. 26, no. 14, pp. 1783–1785, 2010.

[81] M. P. Cox, D. A. Peterson, and P. J. Biggs, “SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data,” *BMC bioinformatics*, vol. 11, no. 1, p. 485, 2010.

[82] E. Planet, C. S.-O. Attolini, O. Reina, O. Flores, and D. Rossell, “htSeqTools: high-throughput sequencing quality control, processing and visualization in R,” *Bioinformatics*, vol. 28, no. 4, pp. 589–590, 2012.

[83] H. Li and R. Durbin, “Fast and accurate long-read alignment with Burrows–Wheeler transform,” *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[84] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *et al.*, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biology*, vol. 10, no. 3, p. R25, 2009.

[85] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang, “SOAP2: an improved ultrafast tool for short read alignment,” *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[86] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[87] Z. Ning, A. J. Cox, and J. C. Mullikin, “SSAHA: a fast search method for large DNA databases,” *Genome Research*, vol. 11, no. 10, pp. 1725–1729, 2001.

[88] V. L. Galinsky, “YOABS: Yet other aligner of biological sequencesan efficient linearly scaling nucleotide aligner,” *Bioinformatics*, vol. 28, no. 8, pp. 1070–1077, 2012.

[89] H. Li and N. Homer, “A survey of sequence alignment algorithms for next-generation sequencing,” *Briefings in Bioinformatics*, vol. 11.

[90] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, *et al.*, “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” *Nature Genetics*, vol. 43, no. 5, pp. 491–498, 2011.

[91] V. Bansal, “A statistical method for the detection of variants from next-generation resequencing of DNA pools,” *Bioinformatics*, vol. 26, no. 12, pp. i318–i324, 2010.

[92] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, *et al.*, “The sequence alignment/map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[93] Z. Wei, W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson, “SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data,” *Nucleic Acids Research*, vol. 39, no. 19, pp. e132–e132, 2011.

[94] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, pp. e164–e164, 2010.

[95] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O'Donnell-Luria, J. Ware, A. Hill, B. Cummings, *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *BioRxiv*, p. 030338, 2016.

[96] G. M. Spudich and X. M. Fernández-Suárez, "Touring Ensembl: a practical guide to genome browsing," *BMC Genomics*, vol. 11, no. 1, p. 295, 2010.

[97] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, *et al.*, "The UCSC Genome Browser database: extensions and updates 2013," *Nucleic Acids Research*, vol. 41, no. D1, pp. D64–D69, 2013.

[98] T. Carver, S. R. Harris, M. Berriman, J. Parkhill, and J. A. McQuillan, "Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data," *Bioinformatics*, vol. 28, no. 4, pp. 464–469, 2012.

[99] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, p. bbs017, 2012.

[100] M. Fiume, V. Williams, A. Brook, and M. Brudno, "Savant: genome browser for high-throughput sequencing data," *Bioinformatics*, vol. 26, no. 16, pp. 1938–1944, 2010.

[101] T. Carver, S. R. Harris, T. D. Otto, M. Berriman, J. Parkhill, and J. A. McQuillan, "BamView: visualizing and interpretation of next-generation sequencing read alignments," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 203–212, 2013.

[102] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[103] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[104] A. Siepel, K. S. Pollard, and D. Haussler, "New methods for detecting lineage-specific selection," in *Research in Computational Molecular Biology*, pp. 190–205, Springer, 2006.

[105] G. M. Cooper, D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad, J. Shendure, and D. A. Nickerson, "Single-nucleotide evolutionary constraint scores highlight disease-causing mutations," *Nature Methods*, vol. 7, no. 4, pp. 250–251, 2010.

[106] L. E. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, *et al.*, "A de novo paradigm for mental retardation," *Nature Genetics*, vol. 42, no. 12, pp. 1109–1112, 2010.

[107] D. MacArthur, T. Manolio, D. Dimmock, H. Rehm, J. Shendure, G. Abecasis, D. Adams, R. Altman, S. Antonarakis, E. Ashley, *et al.*, "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–476, 2014.

[108] C. J. Mattocks, M. A. Morris, G. Matthijs, E. Swinnen, A. Corveleyn, E. Dequeker, C. R. Müller, V. Pratt, and A. Wallace, “A standardized framework for the validation and verification of clinical molecular genetic tests,” *European Journal of Human Genetics*, vol. 18, no. 12, pp. 1276–1288, 2010.

[109] A. S. Gargis, L. Kalman, M. W. Berry, D. P. Bick, D. P. Dimmock, T. Hambuch, F. Lu, E. Lyon, K. V. Voelkerding, B. A. Zehnbauer, *et al.*, “Assuring the quality of next-generation sequencing in clinical laboratory practice,” *Nature Biotechnology*, vol. 30, no. 11, pp. 1033–1036, 2012.

[110] N. Aziz, Q. Zhao, L. Bry, D. K. Driscoll, B. Funke, J. S. Gibson, W. W. Grody, M. R. Hegde, G. A. Hoeltge, D. Leonard, *et al.*, “College of american pathologists’ laboratory standards for next-generation sequencing clinical tests.,” *Archives of Pathology & Laboratory Medicine*, vol. 139, no. 4, pp. 481–493, 2015.

[111] I. Schrijver, N. Aziz, D. H. Farkas, M. Furtado, A. F. Gonzalez, T. C. Greiner, W. W. Grody, T. Hambuch, L. Kalman, J. A. Kant, *et al.*, “Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology,” *The Journal of Molecular Diagnostics*, vol. 14, no. 6, pp. 525–540, 2012.

[112] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, “Statistical challenges associated with detecting copy number variations with next-generation sequencing,” *Bioinformatics*, vol. 28, no. 21, pp. 2711–2718, 2012.

[113] V. Thakur and R. Varshney, “Challenges and strategies for next generation sequencing (NGS) data analysis,” *Journal of Computer Science & Systems Biology*, vol. 3, pp. 40–42, 2010.

[114] R. Altman, “Direct-to-consumer genetic testing: failure is not an option,” *Clinical Pharmacology & Therapeutics*, vol. 86, no. 1, pp. 15–17, 2009.

[115] T. Caulfield, N. Ries, P. Ray, C. Shuman, and B. Wilson, “Direct-to-consumer genetic testing: good, bad or benign?,” *Clinical Genetics*, vol. 77, no. 2, pp. 101–105, 2010.

[116] L. Edge, “Direct-to-consumer genetic tests: flawed and unethical,” *Lancet Oncology*, vol. 9, p. 1113, 2008.

[117] J. K. Teer and J. C. Mullikin, “Exome sequencing: the sweet spot before whole genomes,” *Human Molecular Genetics*, p. ddq333, 2010.

[118] S. M. Group, “Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases,” *Genome Biology*, vol. 16, no. 1, 2015.

[119] C.-S. Ku, D. N. Cooper, C. Polychronakos, N. Naidoo, M. Wu, and R. Soong, “Exome sequencing: dual role as a discovery and diagnostic tool,” *Annals of neurology*, vol. 71, no. 1, pp. 5–14, 2012.

[120] J. Altmüller, B. S. Budde, and P. Nürnberg, “Enrichment of target sequences for next-generation sequencing applications in research and diagnostics,” *Biological Chemistry*, vol. 395, no. 2.

[121] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, E. Lyon, *et al.*, “ACMG clinical laboratory standards for next-generation sequencing,” *Genetics in Medicine*, vol. 15, no. 9, pp. 733–747, 2013.

[122] S. R. Head, H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, and P. Ordoukhanian, “Library construction for next-generation sequencing: overviews and challenges,” *Biotechniques*, vol. 56, no. 2, p. 61, 2014.

[123] L. D. Stein *et al.*, “The case for cloud computing in genome informatics,” *Genome Biology*, vol. 11, no. 5, p. 207, 2010.

[124] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, *et al.*, “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” *Genetics in Medicine*, 2015.

[125] R. C. Green, J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, R. L. Nussbaum, J. M. ODaniel, K. E. Ormond, *et al.*, “ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing,” *Genetics in Medicine*, vol. 15, no. 7, pp. 565–574, 2013.

[126] J. Thusberg, A. Olatubosun, and M. Vihinen, “Performance of mutation pathogenicity prediction methods on missense variants,” *Human Mutation*, vol. 32, no. 4, pp. 358–368, 2011.

[127] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, “Predicting the functional effect of amino acid substitutions and indels,” *PloS one*, vol. 7, no. 10, 2012.

[128] H. G. V. Society *et al.*, “Nomenclature for the description of sequence variants,” 2014.

[129] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, “The next-generation sequencing revolution and its impact on genomics,” *Cell*, vol. 155, no. 1, pp. 27–38, 2013.

[130] E. Commission, “European Commission Regulation (EC) No 141/2000 of the European Parliament and the Council of 16 december 1999 on orphan medicinal products,” *Official Journal of the European Communities*, vol. L18 (1), 2000.

[131] N. II, “Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions,” 2011.

[132] T. Costa, C. R. Scriver, B. Childs, J. M. Opitz, and J. F. Reynolds, “The effect of Mendelian disease on human health: a measurement,” *American Journal of Medical Genetics*, vol. 21, no. 2, pp. 231–242, 1985.

[133] K. D. Farwell, L. Shahmirzadi, D. El-Khechen, Z. Powis, E. C. Chao, B. T. Davis, R. M. Baxter, W. Zeng, C. Mroske, M. C. Parra, *et al.*, “Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions,” *Genetics in Medicine*, 2014.

[134] P. Wangkumhang, K. Chaichoompu, C. Ngamphiw, U. Ruangrit, J. Chanprasert, A. Assawamakin, and S. Tongsima, “WASP: a Web-based Allele-Specific PCR assay designing tool for detecting snps and mutations,” *BMC Genomics*, vol. 8, no. 1, p. 275, 2007.

[135] W. Moore and M. Guille, "Preparation and testing of synthetic mRNA for microinjection," in *Molecular Methods in Developmental Biology*, pp. 99–109, Springer, 1999.

[136] S. Bione, F. Rizzolio, C. Sala, R. Ricotti, M. Goegan, M. Manzini, R. Battaglia, A. Marozzi, W. Vegetti, L. Dalpra, *et al.*, "Mutation analysis of two candidate genes for premature ovarian failure, DACH2 and POF1B," *Human Reproduction*, vol. 19, no. 12, pp. 2759–2766, 2004.

[137] W. Fassnacht, A. Mempel, T. Strowitzki, and P. Vogt, "Premature ovarian failure (POF) syndrome: towards the molecular clinical analysis of its genetic complexity," *Current Medicinal Chemistry*, vol. 13, no. 12, pp. 1397–1410, 2006.

[138] A. Gijsbers, N. Den Hollander, A. Helderman-van de Enden, J. Schuurs-Hoeijmakers, L. Vijfhuizen, E. Bijlsma, A. Van Haeringen, K. Hansson, E. Bakker, M. Breuning, *et al.*, "X-chromosome duplications in males with mental retardation: pathogenic or benign variants?," *Clinical Genetics*, vol. 79, no. 1, pp. 71–78, 2011.

[139] A. J. Stevens, H. L. Kennedy, and M. A. Kennedy, "Fluorescent methods for probing G-quadruplex structure in single and double stranded dna," *Biochemistry*, 2016.

[140] M. Ruault, M. Ventura, N. Galtier, M.-E. Brun, N. Archidiacono, G. érard Roizès, and A. De Sario, "BAGE genes generated by juxtacentromeric reshuffling in the hominidae lineage are under selective pressure," *Genomics*, vol. 81, no. 4, pp. 391–399, 2003.

[141] E. S. Hopmans, G. Natsoulis, J. M. Bell, S. M. Grimes, W. Sieh, and H. P. Ji, "A programmable method for massively parallel targeted sequencing," *Nucleic Acids Research*, vol. 42, no. 10, pp. e88–e88, 2014.

[142] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, *et al.*, "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," *Nature*, vol. 423, no. 6942, pp. 825–837, 2003.

[143] M. Johansson, "Cellular sexual dimorphism of X and Y homolog gene expression in human central nervous system during early male development," 2015.

[144] M. Fukami, S. Kirsch, S. Schiller, A. Richter, V. Benes, B. Franco, K. Muroya, E. Rao, S. Merker, B. Niesler, *et al.*, "A member of a gene family on Xp22. 3, VCX-A, is deleted in patients with X-linked nonspecific mental retardation," *The American Journal of Human Genetics*, vol. 67, no. 3, pp. 563–573, 2000.

[145] C. Tyson, A. Dawson, S. Bal, M. Tomiuk, T. Anderson, D. Tucker, D. Riordan, I. Chudoba, B. Morash, A. Mhanni, *et al.*, "Molecular cytogenetic investigation of two patients with Y chromosome rearrangements and intellectual disability," *American Journal of Medical Genetics Part A*, vol. 149, no. 3, pp. 490–495, 2009.

[146] L. Huang, L. A. Jolly, S. Willis-Owen, A. Gardner, R. Kumar, E. Douglas, C. Shoubridge, D. Wieczorek, A. Tzschach, M. Cohen, *et al.*, "A noncoding, regulatory mutation implicates HCFC1 in nonsyndromic intellectual disability," *The American Journal of Human Genetics*, vol. 91, no. 4, pp. 694–702, 2012.

[147] B. J. O'Roak, P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. MacKenzie, S. B. Ng, C. Baker, *et al.*, "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations," *Nature Genetics*, vol. 44, no. 4, pp. 471–471, 2012.

[148] J. de Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. de Vries, C. Gilissen, *et al.*, "Diagnostic exome sequencing in persons with severe intellectual disability," *New England Journal of Medicine*, vol. 367, no. 20, pp. 1921–1929, 2012.

[149] A.-R. Lee, J. Park, K. J. Jung, S. H. Jee, and S. Kim-Yoon, "Genetic variation rs7930 in the mir-4273-5p target site is associated with a risk of colorectal cancer," *OncoTargets and Therapy*, vol. 9, p. 6885, 2016.

[150] S. Razin, V. Borunova, O. Maksimenko, and O. Kantidze, "Cys2His2 zinc finger protein family: classification, functions, and major members," *Biochemistry (Moscow)*, vol. 77, no. 3, pp. 217–226, 2012.

[151] A. G. Baranovskiy, Y. Zhang, Y. Suwa, N. D. Babayeva, J. Gu, Y. I. Pavlov, and T. H. Tahirov, "Crystal structure of the human primase," *Journal of Biological Chemistry*, vol. 290, no. 9, pp. 5635–5646, 2015.

[152] K. D. Huynh, W. Fischle, E. Verdin, and V. J. Bardwell, "BCOR, a novel corepressor involved in BCL-6 repression," *Genes & Development*, vol. 14, no. 14, pp. 1810–1823, 2000.

[153] D. Horn, M. Chyrek, S. Kleier, S. Lüttgen, H. Bolz, G.-K. Hinkel, G. C. Korenke, A. Rieß, C. Schell-Apacik, S. Tinschert, *et al.*, "Novel mutations in BCOR in three patients with Oculo-Facio-Cardio-Dental Syndrome, but none in Lenz microphthalmia syndrome," *European Journal of Human Genetics*, vol. 13, no. 5, pp. 563–569, 2005.

[154] Y. Yamamoto, A. Abe, and N. Emi, "Clarifying the impact of polycomb complex component disruption in human cancers," *Molecular Cancer Research*, vol. 12, no. 4, pp. 479–484, 2014.

[155] D. A. Kleinjan and V. van Heyningen, "Long-range control of gene expression: emerging mechanisms and disruption in disease," *The American Journal of Human Genetics*, vol. 76, no. 1, pp. 8–32, 2005.

[156] D.-J. Kleinjan and P. Coutinho, "Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease," *Briefings in Functional Genomics & Proteomics*, vol. 8, no. 4, pp. 317–332, 2009.

[157] R. J. Gorlin, A. H. Marashi, and H. L. Obwegeser, "Oculo-Facio-Cardio-Dental (OFCD) syndrome," *European Journal of Human Genetics*, vol. 63, no. 1, pp. 290–292, 1996.

[158] D. Ng, N. Thakker, C. M. Corcoran, D. Donnai, R. Perveen, A. Schneider, D. W. Hadley, C. Tifft, L. Zhang, A. O. Wilkie, *et al.*, "Oculofaciocardiodental and Lenz microphthalmia syndromes result from distinct classes of mutations in BCOR," *Nature Genetics*, vol. 36, no. 4, pp. 411–416, 2004.

[159] E. Hilton, J. Johnston, S. Whalen, N. Okamoto, Y. Hatsukawa, J. Nishio, H. Kohara, Y. Hirano, S. Mizuno, C. Torii, *et al.*, "BCOR analysis in patients with OFCD and Lenz microphthalmia syndromes, mental retardation with ocular anomalies, and cardiac laterality defects," *European Journal of Human Genetics*, vol. 17, no. 10, pp. 1325–1335, 2009.

[160] J. A. Wamstad, C. M. Corcoran, A. M. Keating, and V. J. Bardwell, "Role of the transcriptional corepressor bcor in embryonic stem cell differentiation and early embryonic development," *PLoS One*, vol. 3, no. 7, p. e2814, 2008.

[161] Z. Fan, T. Yamaza, J. S. Lee, J. Yu, S. Wang, G. Fan, S. Shi, and C.-Y. Wang, "BCOR regulates mesenchymal stem cell function by epigenetic mechanisms," *Nature Cell Biology*, vol. 11, no. 8, pp. 1002–1009, 2009.

[162] T. Surapornsawasd, T. Ogawa, and K. Moriyama, "Identification of nuclear localization signals within the human BCOR protein," *FEBS letters*, vol. 589, no. 21, pp. 3313–3320, 2015.

[163] P. N. Kantaputra, "BCOR mutations and unstoppable root growth: a commentary on oculofaciocardiodental syndrome: novel BCOR mutations and expression in dental cells," *Journal of Human Genetics*, vol. 59, no. 6, pp. 297–299, 2014.

[164] T. Surapornsawasd, T. Ogawa, M. Tsuji, and K. Moriyama, "Oculofaciocardiodental syndrome: novel BCOR mutations and expression in dental cells," *Journal of Human Genetics*, vol. 59, no. 6, pp. 314–320, 2014.

[165] V. Grossmann, E. Tiacci, A. B. Holmes, A. Kohlmann, M. P. Martelli, W. Kern, A. Spanhol-Rosseto, H.-U. Klein, M. Dugas, S. Schindela, *et al.*, "Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype," *Blood*, vol. 118, no. 23, pp. 6153–6163, 2011.

[166] F. Damm, V. Chesnais, Y. Nagata, K. Yoshida, L. Scourzic, Y. Okuno, R. Itzykson, M. Sanada, Y. Shiraishi, V. Gelsi-Boyer, *et al.*, "BCOR and BCORL1 mutations in myelodysplastic syndromes and related disorders," *Blood*, vol. 122, no. 18, pp. 3169–3177, 2013.

[167] T. J. Pugh, S. D. Weeraratne, T. C. Archer, D. A. P. Krummel, D. Auclair, J. Bochicchio, M. O. Carneiro, S. L. Carter, K. Cibulskis, R. L. Erlich, *et al.*, "Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations," *Nature*, vol. 488, no. 7409, pp. 106–110, 2012.

[168] J. Zhang, C. A. Benavente, J. McEvoy, J. Flores-Otero, L. Ding, X. Chen, A. Ulyanov, G. Wu, M. Wilson, J. Wang, *et al.*, "A novel retinoblastoma therapy from genomic and epigenetic analyses," *Nature*, vol. 481, no. 7381, pp. 329–334, 2012.

[169] L. Carrel and H. F. Willard, "X-inactivation profile reveals extensive variability in X-linked gene expression in females," *Nature*, vol. 434, no. 7031, pp. 400–404, 2005.

[170] C. Hatakeyama, C. Anderson, C. Beever, M. Penaherrera, C. Brown, and W. Robinson, "The dynamics of x-inactivation skewing as women age," *Clinical Genetics*, vol. 66, no. 4, pp. 327–332, 2004.

[171] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, *et al.*, “Rate of de novo mutations and the importance of father/’s age to disease risk,” *Nature*, vol. 488, no. 7412, pp. 471–475, 2012.

[172] E. Moustacchi, “Fanconis anemia,” *Update*, 2003.

[173] T. Otsuki, J. Wang, I. Demuth, M. Digweed, and J. M. Liu, “Assessment of mitomycin C sensitivity in Fanconi anemia complementation group C gene (Fac) knock-out mouse cells.,” *International Journal of Hematology*, vol. 67, no. 3, pp. 243–248, 1998.

[174] M. Castella, R. Pujol, E. Callén, M. J. Ramírez, J. A. Casado, M. Talavera, T. Ferro, A. Muñoz, J. Sevilla, L. Madero, *et al.*, “Chromosome fragility in patients with Fanconi anaemia: diagnostic implications and clinical impact,” *Journal of Medical Genetics*, vol. 48, no. 4, pp. 242–250, 2011.

[175] S. Danda, V. A. van Rahden, D. John, P. Paul, R. Raju, S. Koshy, and K. Kutsche, “Evidence of germline mosaicism for a novel BCOR mutation in two indian sisters with Oculo-Facio-Cardio-Dental Syndrome,” *Molecular Syndromology*, vol. 5, no. 5, p. 251, 2014.

[176] J. Soulier, “Fanconi anemia,” *ASH Education Program Book*, vol. 2011, no. 1, pp. 492–497, 2011.

[177] J. A. Veltman and H. G. Brunner, “De novo mutations in human genetic disease,” *Nature Reviews Genetics*, vol. 13, no. 8, pp. 565–575, 2012.

[178] D. F. Conrad, J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, *et al.*, “Variation in genome-wide mutation rates within and between human families,” *Nature*, vol. 201, no. 1, 2011.

[179] J. F. Crow, “The origins, patterns and implications of human spontaneous mutation,” *Nature Reviews Genetics*, vol. 1, no. 1, pp. 40–47, 2000.

[180] A. Eyre-Walker and P. D. Keightley, “The distribution of fitness effects of new mutations,” *Nature Reviews Genetics*, vol. 8, no. 8, pp. 610–618, 2007.

[181] M. Lynch, “Rate, molecular spectrum, and consequences of human mutation,” *Proceedings of the National Academy of Sciences*, p. 200912629, 2010.

[182] A. Itsara, H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, S. J. London, and E. E. Eichler, “De novo rates and selection of large copy number variation,” *Genome Research*, vol. 20, no. 11, pp. 1469–1481, 2010.

[183] S. Aretz, D. Stienen, N. Friedrichs, S. Stemmler, S. Uhlhaas, N. Rahner, P. Propping, and W. Friedl, “Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP),” *Human Mutation*, vol. 28, no. 10, pp. 985–992, 2007.

[184] A. Goriely, H. Lord, J. Lim, D. Johnson, T. Lester, H. V. Firth, and A. O. Wilkie, “Germline and somatic mosaicism for FGFR2 mutation in the mother of a child with crouzon syndrome: Implications for genetic testing in paternal age-effect syndromes,” *American Journal of Medical Genetics Part A*, vol. 152, no. 8, pp. 2067–2073, 2010.

[185] R. Acuna-Hidalgo, T. Bo, M. P. Kwint, M. van de Vorst, M. Pinelli, J. A. Veltman, A. Hoischen, L. E. Vissers, and C. Gilissen, "Post-zygotic point mutations are an underrecognized source of De Novo genomic variation," *The American Journal of Human Genetics*, vol. 97, no. 1, pp. 67–74, 2015.

[186] I. M. Campbell, B. Yuan, C. Robberecht, R. Pfundt, P. Szafranski, M. E. McEntagart, S. C. Nagamani, A. Erez, M. Bartnik, B. Wiśniowiecka-Kowalnik, *et al.*, "Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders," *The American Journal of Human Genetics*, vol. 95, no. 2, pp. 173–182, 2014.

[187] M. D. Gearhart, C. M. Corcoran, J. A. Wamstad, and V. J. Bardwell, "Polycomb group and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to BCL6 targets," *Molecular and Cellular Biology*, vol. 26, no. 18, pp. 6880–6889, 2006.

[188] B. R. Schulze, D. Horn, A. Kobelt, G. Tariverdian, and A. Stellzig, "Rare dental abnormalities seen in oculo-facio-cardio-dental (OFCD) syndrome: three new cases and review of nine patients," *American journal of medical genetics*, vol. 82, no. 5, pp. 429–435, 1999.

[189] E. N. Hilton, F. D. Manson, J. E. Urquhart, J. J. Johnston, A. M. Slavotinek, P. Hedera, E.-L. Stattin, A. Nordgren, L. G. Biesecker, and G. C. Black, "Left-sided embryonic expression of the BCL-6 corepressor, BCOR, is required for vertebrate laterality determination," *Human Molecular Genetics*, vol. 16, no. 14, pp. 1773–1782, 2007.

[190] P. Sicinski, J. L. Donaher, Y. Geng, S. B. Parker, H. Gardner, M. Y. Park, R. L. Robker, J. S. Richards, L. K. McGinnis, J. D. Biggers, *et al.*, "Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis," 1996.

[191] M. Greenblatt, W. Bennett, M. Hollstein, and C. Harris, "Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis," *Cancer research*, vol. 54, no. 18, pp. 4855–4878, 1994.

[192] A. L. Dent, A. L. Shaffer, X. Yu, D. Allman, and L. M. Staudt, "Control of inflammation, cytokine expression, and germinal center formation by BCL-6," *Science*, vol. 276, no. 5312, pp. 589–592, 1997.

[193] H. Y. Bihui, G. Cattorettil, Q. Shenl, I. Zhangl, N. Hawe, and R. de Waard, "The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type," *Nature Genetics*, vol. 16, p. 161, 1997.

[194] L. Busque, R. Mio, J. Mattioli, E. Brais, N. Blais, Y. Lalonde, M. Maragh, and D. G. Gilliland, "Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age," *Blood*, vol. 88, no. 1, pp. 59–65, 1996.

[195] P. Shyamsunder, K. S. Ganesh, P. Vidyasekar, S. Mohan, and R. S. Verma, "Identification of novel target genes involved in Indian Fanconi anemia patients using microarray," *Gene*, vol. 531, no. 2, pp. 444–450, 2013.

[196] N. Norton, D. Li, E. Rampersaud, A. Morales, E. R. Martin, S. Zuchner, S. Guo, M. Gonzalez, D. J. Hedges, P. D. Robertson, *et al.*, "Exome sequencing and genome-wide linkage analysis in 17 families illustrates the complex contribution of TTN truncating variants to dilated cardiomyopathy," *Circulation: Cardiovascular Genetics*, pp. CIRCGENETICS–113, 2013.

[197] S. Curbo, C. Lagier-Tourenne, R. Carrozzo, L. Palenzuela, S. Lucioli, M. Hirano, F. Santorelli, J. Arenas, A. Karlsson, and M. Johansson, "Human mitochondrial pyrophosphatase: cDNA cloning and analysis of the gene in patients with mtDNA depletion syndromes," *Genomics*, vol. 87, no. 3, pp. 410–416, 2006.

[198] S. Cohen, R. Sterner, P. Keim, and R. Heinrikson, "Covalent structural analysis of yeast inorganic pyrophosphatase.," *Journal of Biological Chemistry*, vol. 253, no. 3, pp. 889–897, 1978.

[199] M. Lundin, H. Baltscheffsky, and H. Ronne, "Yeast PPA2 gene encodes a mitochondrial inorganic pyrophosphatase that is essential for mitochondrial function.," *Journal of Biological Chemistry*, vol. 266, no. 19, pp. 12168–12172, 1991.

[200] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[201] Y.-C. Shen, J.-H. Fan, H. J. Edenberg, T.-K. Li, Y.-H. Cui, Y.-F. Wang, C.-H. Tian, C.-F. Zhou, R.-L. Zhou, J. Wang, *et al.*, "Polymorphism of ADH and ALDH genes among four ethnic groups in China and effects upon the risk for alcoholism," *Alcoholism: Clinical and Experimental Research*, vol. 21, no. 7, pp. 1272–1277, 1997.

[202] D. C. Wallace, "Mitochondrial defects in cardiomyopathy and neuromuscular disease," *American Heart Journal*, vol. 139, no. 2, pp. s70–s85, 2000.

[203] C. M. Sue and E. A. Schon, "Mitochondrial respiratory chain diseases and mutations in nuclear DNA: a promising start?," *Brain Pathology*, vol. 10, no. 3, pp. 442–450, 2000.

[204] A. Barrientos, F. Fontanesi, and F. Díaz, "Evaluation of the mitochondrial respiratory chain and oxidative phosphorylation system using polarography and spectrophotometric enzyme assays," *Current Protocols in Human Genetics*, pp. 19–3, 2009.

[205] M. Alexeyev, I. Shokolenko, G. Wilson, and S. LeDoux, "The maintenance of mitochondrial DNA integrity. critical analysis and update," *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 5, p. a012641, 2013.

[206] J. Leonard and A. H. Schapira, "Mitochondrial respiratory chain disorders i: mitochondrial DNA defects," *The Lancet*, vol. 355, no. 9200, pp. 299–304, 2000.

[207] D. Thorburn, "Mitochondrial disorders: prevalence, myths and advances," *Journal of Inherited Metabolic Disease*, vol. 27, no. 3, pp. 349–362, 2004.

[208] D. C. Wallace, "Mitochondrial diseases in man and mouse," *Science*, vol. 283, no. 5407, pp. 1482–1488, 1999.

[209] K. A. Strauss, L. DuBiner, M. Simon, M. Zaragoza, P. P. Sengupta, P. Li, N. Narula, S. Dreike, J. Platt, V. Procaccio, *et al.*, "Severity of cardiomyopathy associated with adenine nucleotide translocator-1 deficiency correlates with mtDNA haplogroup," *Proceedings of the National Academy of Sciences*, vol. 110, no. 9, pp. 3453–3458, 2013.

[210] M. G. Castro, C. Huerta, J. R. Reguero, M. I. Soto, E. Doménech, V. Alvarez, M. Gómez-Zaera, V. Nunes, P. González, A. Corao, *et al.*, "Mitochondrial DNA haplogroups in

Spanish patients with hypertrophic cardiomyopathy," *International Journal of Cardiology*, vol. 112, no. 2, pp. 202–206, 2006.

[211] F. Palmieri and C. L. Pierri, "Mitochondrial metabolite transport," *Essays in Biochemistry*, vol. 47, pp. 37–52, 2010.

[212] E. R. Kunji, "The role and structure of mitochondrial carriers," *FEBS letters*, vol. 564, no. 3, pp. 239–244, 2004.

[213] R. Kramer, "Characterization of pyrophosphate exchange by the reconstituted adenine nucleotide translocator from mitochondria," *Biochemical and Biophysical Research Communications*, vol. 127, no. 1, pp. 129–135, 1985.

[214] H. Aquila, T. A. Link, and M. Klingenberg, "Solute carriers involved in energy transfer of mitochondria form a homologous protein family," *FEBS letters*, vol. 212, no. 1, pp. 1–9, 1987.

[215] F. Palmieri, "Mitochondrial transporters of the SLC25 family and associated diseases: a review," *Journal of Inherited Metabolic Disease*, vol. 37, no. 4, pp. 565–575, 2014.

[216] A. Götz, H. Tyynismaa, L. Euro, P. Ellonen, T. Hyötyläinen, T. Ojala, R. H. Hämäläinen, J. Tommiska, T. Raivio, M. Oresic, *et al.*, "Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy," *The American Journal of Human Genetics*, vol. 88, no. 5, pp. 635–642, 2011.

[217] J. W. Yarham, T. N. Lamichhane, A. Pyle, S. Mattijssen, E. Baruffini, F. Bruni, C. Donnini, A. Vassilev, L. He, E. L. Blakely, *et al.*, "Defective i 6 A37 modification of mitochondrial and cytosolic tRNAs results from pathogenic mutations in TRIT1 and its substrate tRNA," *PLoS Genetics*, vol. 10, no. 6, p. e1004424, 2014.

[218] A. M. Davidson and A. P. Halestrap, "Liver mitochondrial pyrophosphate concentration is increased by $Ca2^{+}$ and regulates the intramitochondrial volume and adenine nucleotide content.," *Biochemical Journal*, vol. 246, pp. 715–723, 1987.

[219] L. F. C. Castro, M. Lopes-Marques, J. M. Wilson, E. Rocha, M. A. Reis-Henriques, M. M. Santos, and I. Cunha, "A novel Acetyl-CoA synthetase short-chain subfamily member 1 (Acss1) gene indicates a dynamic history of paralogue retention and loss in vertebrates," *Gene*, vol. 497, no. 2, pp. 249–255, 2012.

[220] S. Zakhari, "Overview: how is alcohol metabolized by the body?," *Alcohol Research & Health*, 2006.

[221] M. J. Ackerman, D. J. Tester, and C.-B. J. Porter, "Swimming, a gene-specific arrhythmogenic trigger for inherited long QT syndrome," in *Mayo Clinic Proceedings*, vol. 74, pp. 1088–1094, Elsevier, 1999.

[222] P. J. Schwartz, S. G. Priori, C. Spazzolini, A. J. Moss, G. M. Vincent, C. Napolitano, I. Denjoy, P. Guicheney, G. Breithardt, M. T. Keating, *et al.*, "Genotype-phenotype correlation in the long-QT syndrome gene-specific triggers for life-threatening arrhythmias," *Circulation*, vol. 103, no. 1, pp. 89–95, 2001.

[223] T. Nakajima, K. Misu, K. Iwasawa, E. Tamiya, K. Segawa, H. Matsuo, and K. Hada, "Auditory stimuli as a major cause of syncope in a patient with idiopathic long QT syndrome.," *Japanese Circulation Journal*, vol. 59, no. 4, pp. 241–246, 1995.

[224] A. P. Halestrap, "A pore way to die: the role of mitochondria in reperfusion injury and cardioprotection," *Biochemical Society Transactions*, vol. 38, no. 4, pp. 841–860, 2010.

[225] T. E. North and L. I. Zon, "Modeling human hematopoietic and cardiovascular diseases in zebrafish," *Developmental Dynamics*, vol. 228, no. 3, pp. 568–583, 2003.

[226] C. C. Leitch, N. A. Zaghloul, E. E. Davis, C. Stoetzel, A. Diaz-Font, S. Rix, M. Alfadhel, R. A. Lewis, W. Eyaid, E. Banin, *et al.*, "Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome," *Nature Genetics*, vol. 40, no. 4, pp. 443–448, 2008.

[227] N. N. Aung, H. Kennedy, J. M. Faed, and S. O. Brennan, "Novel heterozygous Bbeta (c. 1311T>A) mutation (Fibrinogen St Kilda) associated with recurrent pregnancy loss," *Pathology-Journal of the RCPA*, vol. 47, no. 6, pp. 583–585, 2015.

[228] G. Kularatnam, D. Warawita, S. Jayasena, S. Nadarajah, E. Jasinge, D. Mendis, H. Kennedy, C. Florkowski, and P. George, "Cystic fibrosis in a Sri Lankan infant, confirmed by genotyping: implications for future diagnosis and service provision," *Journal of the Postgraduate Institute of Medicine*, vol. 2, 2015.