APPLICATION OF CLASSICAL AND NEXT GENERATION SEQUENCING TECHNIQUES TO GENE DISCOVERY AND DIAGNOSTICS IN NEUROGENETIC DISORDERS

Kyle Yau
BSc (Molecular Biology), Grad Dip Sci (Physiology)

This thesis is presented for the degree of
Doctor of Philosophy
Of The University of Western Australia

School of Medicine and Pharmacology

2014

THE UNIVERSITY OF
WESTERN AUSTRALIA

DECLARATION

This thesis contains published work and/or work prepared for publication, some of which has been co-authored. The bibliographical details of the work and where it appears in the thesis are outlined below. All work presented in the thesis was conducted by the candidate unless otherwise indicated in the text.

Chapter Two
Sambuughin N, **Yau KS**, Olivé M, Duff RM, Bayarsaikhan M, Lu S, Gonzalez-Mera L, Sivadorai P, Nowak KJ, Ravenscroft G, Mastaglia FL, North KN, Ilkovski B, Kremer H, Lammens M, van Engelen BG, Fabian V, Lamont P, Davis MR, Laing NG, Goldfarb LG. (2010) Dominant mutations in KBTBD13, a member of the BTB/Kelch family, cause nemaline myopathy with cores. *American Journal of Human Genetics* 87(6):842-7
Contribution by candidate: 30%

Yau, K. S., Olivé, M., Lamont, P. J., & Laing, N. G. (2013) Kelch Proteins. *Muscle Disease: Pathology and Genetics, Second edition*, 252-253.
Contribution by candidate: 80%

Chapter Four
Ravenscroft G, Thompson EM, Todd EJ, **Yau KS**, Kresoje N, Sivadorai P, Friend K, Riley K, Manton ND, Blumbergs P, Fietz M, Duff RM, Davis MR, Allcock RJ, Laing NG. (2013) Whole exome sequencing in foetal akinesia expands the genotype-phenotype spectrum of GBE1 glycogen storage disease mutations. *Neuromuscular Disorders* 23(2):165-9
Contribution by candidate: 15%

Ravenscroft G, Miyatake S, Lehtokari VL, Todd EJ, Vornanen P, **Yau KS**, Hayashi YK, Miyake N, Tsurusaki Y, Doi H, Saitsu H, Osaka H, Yamashita S, Ohya T, Sakamoto Y, Koshimizu E, Imamura S, Yamashita M, Ogata K, Shiina M, Bryson-Richardson RJ, Vaz R, Ceyhan O, Brownstein CA, Swanson LC, Monnot S, Romero NB, Amthor H, Kresoje N, Sivadorai P, Kiraly-Borri C, Haliloglu G, Talim B, Orhan D, Kale G, Charles AK, Fabian VA, Davis MR, Lammens M, Sewry CA, Manzur A, Muntoni F, Clarke NF, North KN, Bertini E, Nevo Y, Willichowski E, Silberg IE, Topaloglu H, Beggs AH, Allcock RJ, Nishino I, Wallgren-Pettersson C, Matsumoto N, Laing NG. (2013) Mutations in KLHL40 are a frequent cause of severe autosomal-recessive nemaline myopathy. *American Journal of Human Genetics* 93(1):6-18
Contribution by candidate: 10%

Gupta VA, Ravenscroft G, Shaheen R, Todd EJ, Swanson LC, Shiina M, Ogata K, Hsu C, Clarke NF, Darras BT, Farrar MA, Hashem A, Manton ND, Muntoni F, North KN, Sandaradura SA, Nishino I, Hayashi YK, Sewry CA, Thompson EM, **Yau KS**, Brownstein CA, Yu TW, Allcock RJ, Davis MR, Wallgren-Pettersson C, Matsumoto N, Alkuraya FS, Laing NG, Beggs AH. (2013) Identification of KLHL41 Mutations Implicates BTB-Kelch-Mediated Ubiquitination as an Alternate Pathway to Myofibrillar Disruption in Nemaline Myopathy. *American Journal of Human Genetics* 93(6):1108-17
Contribution by candidate: 5%

Ong RW, AlSaman A, Selcen D, Arabshahi A, **Yau KS**, Ravenscroft G, Duff RM, Atkinson V, Allcock RJ, Laing NG. (2014) Novel cofilin-2 (CFL2) four base pair deletion causing nemaline myopathy. *Journal of neurology, neurosurgery and psychiatry* 85(9):1058-60
Contribution by candidate: 10%

Agrawal PB, Pierson CR, Joshi M, Liu X, Ravenscroft G, Moghadaszadeh B, Talabere T, Viola M, Swanson LC, Haliloğlu G, Talim B, **Yau KS**, Allcock RJ, Laing NG, Perrella MA, Beggs AH. (2014) SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *American Journal of Human Genetics 95(2):218-26*
Contribution by candidate: 5%

Signature……
Kyle Yau – Candidate

Signature……
W/Prof Nigel G Laing – Coordinating Supervisor

PUBLISHED CONFERENCE ABSTRACTS

K.S. Yau, P.J. Lamont, V. Fabian, N. Kresoje, P. Sivadorai, R. Allcock, M. Davis, N.G. Laing (2012) Study of a novel autosomal recessive minicore disease. In 17th International Congress of the World Muscle Society, pp.842 *Neuromuscular Disorders,* Perth
Contribution by candidate: 80%

K.S. Yau, P.J. Lamont, K.J. Nowak, N. Kresoje, E.L. McNamara, R.J. Allcock, M.R. Davis, N.G. Laing (2012) Study of an autosomal recessive spinocerebellar ataxia with peripheral neuropathy. In 17th International Congress of the World Muscle Society, pp.869-870 *Neuromuscular Disorders,* Perth
Contribution by candidate: 80%

P.J. Lamont, **K.S. Yau**, R.M. Duff, W. Carroll, M.R. Davis, N.G. Laing (2012) Study of a novel autosomal dominant spinocerebellar ataxia. In 17th International Congress of the World Muscle Society, pp.870 *Neuromuscular Disorders,* Perth
Contribution by candidate: 40%

G. Ravenscroft, E. Todd, **K.S. Yau**, N. Kresoje, P. Sivadorai, E. Sollis, K. Friend, K. Riley, E. Thompson, N. Manton, P. Blumbergs, C. Kiraly-Borri, G. Haliloglu, D. Orhan, G. Kale, A.K. Charles, V. Fabian, M.R. Davis, R.J. Allcock, N.G. Laing (2012) Whole exome sequencing applied to foetal akinesia. In 17th International Congress of the World Muscle Society, pp.809 *Neuromuscular Disorders,* Perth
Contribution by candidate: 10%

R.W. Ong, P.J. Lamont, **K.S. Yau**, N. Kresoje, M.R. Davis, R.J. Allcock, N.G. Laing Whole exome sequencing applied to Charcot–Marie–Tooth (CMT) disease (2012) In 17th International Congress of the World Muscle Society, pp.808-809 *Neuromuscular Disorders,* Perth
Contribution by candidate: 10%

K. Yau, R. Allcock, K. Mina, G. Ravenscroft, M. Cabrera, R. Gooding, C. Wise, P. Sivadorai, D. Trajanoski, V. Atkinson, S. Wagner, K. Nowak, R. Duff, P. Lamont, M. Davis, N. Laing (2014) Neurogenetic disease diagnostics by targeted capture and next generation sequencing. In 19th International Congress of the World Muscle Society, pp.799-800 *Neuromuscular Disorders*, Berlin
Contribution by candidate: 60%

E.C. Oates, **K.S. Yau**, A. Charlton, S. Brammah, M.A. Farrar, H. Sampai, P.L. Lamont, D. Mowat, R.B. Fitzsimons, A. Corbett, M.M. Ryan, H.L. Teoh, G.L. O'Grady, R. Ghaoui, S. Kaur, M. Lek, K.N. North, D.G. MacArthur, M.R. Davis, N.G. Laing, N.F. Clarke (2014) Analysis of a large patient cohort with recessive truncating TTN mutations reveals novel clinical features and a diverse range of muscle pathologies. In 19th International Congress of the World Muscle Society, pp.805 *Neuromuscular Disorders*, Berlin
Contribution by candidate: 25%

G. Ravenscroft; E.J. Todd; **K.S. Yau**; C.A. Sewry; C.A. McLean; M.M. Ryan; R.J. Allcock; N.G. Laing (2014) Nemaline myopathy 8 and KLHL40 in diseased and normal skeletal muscle. In 19th International Congress of the World Muscle Society, pp.899 *Neuromuscular Disorders*, Berlin
Contribution by candidate: 10%

Signature……
Kyle Yau – Candidate

Signature……
W/Prof Nigel G Laing – Coordinating Supervisor

ABSTRACT

Neurogenetic diseases are a genetically and phenotypically diverse set of diseases that affect the skeletal muscle, central and peripheral nervous systems. This heterogeneity, combined with the relative rareness of each individual disease type, makes identification of novel disease genes and molecular diagnosis difficult endeavours, resulting in the majority of patients going without a molecular diagnosis.

In this thesis, next generation sequencing as well as classical methods were used to facilitate disease gene discovery and molecular diagnosis in a set of novel disease phenotypes, and next generation sequencing was used in the implementation of a neurogenetic sub-exomic sequencing panel, designed to target 336 neurogenetic and cardiac disease genes for diagnostic purposes.

A single novel disease gene Kelch and BTB-domain containing 13 (*KBTBD13*) was identified using the classical disease gene discovery methods of linkage analysis and positional cloning. Using next generation sequencing, and classical methods, three novel disease genes (*KLHL40*, *KLHL41*, *SPEG*) were identified, and the phenotypic spectrum of four known disease genes (*CFL2*, *ECEL1*, *GBE1*, *TTN*) were expanded. The neurogenetic sub-exomic sequencing panel has been successfully deployed as a front-line diagnostic test in Australasia, with a diagnostic success rate of up to 34%.

The results of this thesis support the current trend in disease gene discovery and molecular diagnosis – that next generation sequencing is being translated from a purely research-based tool into one that can be applied on a routine basis to enhance the health of the population. The pace of disease gene discovery is expected to intensify with the increased availability of next generation sequencing technologies, and the influx of improved sequencing methods.

ACKNOWLEDGEMENTS

This thesis is the culmination of more than four years of work, and would not have been possible without the dedicated support, drive and vision that my supervisors Dr. Phillipa Lamont, W/Prof. Nigel Laing, A/Prof Kristen Nowak and Dr. Mark Davis have provided.

It's been a long road, and there's finally light at the end of the tunnel.

Thank you Gina, for your guidance and instruction when teaching me about cell culture, immunohistochemistry, western blotting, and basically everything to do with characterising proteins. Thank you Rachael, for starting me along the winding, bumpy path known as bioinformatics, and helping me in those early days. Thank you Elyshia, for helping me with my insect cell cultures, and being a deft hand in the lab when I've needed it.

To the rest of the Laing Laboratory; Macarena, Jordan, Royston, Adriana, Klair and Tina, thank you for the support and expertise you've provided over the years, and for being excellent friends.

To the staff at the LotteriesWest State Biomedical Facility Genomics, past and present, Richard, Vanessa, Sarah, and Nina, the majority of the work in this thesis would not have been possible without you. Thank you for supporting me, and for employing me, for that matter!

Last but not least, for all the years you've put up with me, for all the years we've been together, a special thank you to the light of my life, Siobhan.

TABLE OF CONTENTS

DECLARATION ii

PUBLISHED CONFERENCE ABSTRACTS iv

ABSTRACT vi

ACKNOWLEDGEMENTS viii

TABLE OF CONTENTS ix

LIST OF FIGURES AND TABLES xxi

LIST OF ABBREVIATIONS xxv

AIMS 1

Chapter 1 2

Introduction: genetic diseases, methods of disease gene discovery

1.1 Aims 3

1.2 Genetic disease 3

1.2.1 The importance of finding a causative disease gene 4

1.2.1.1 Accurate Diagnosis 6

1.2.1.2 Prognosis 6

1.2.1.3 Prenatal, preimplantation and presymptomatic diagnosis 7

1.2.1.4 Investigation of the pathobiology of genetic disease 9

1.2.1.5 Increasing understanding of tissue biology 9

1.2.1.6 Rational development of therapies 10

1.2.1.7 Development of other therapeutics 11

1.3 Identification of disease genes – A short historical perspective 12

1.3.1 Gene discovery before linkage maps and the Human Genome Project 12

1.3.2 Construction of human linkage maps 13

1.3.3 Exclusion mapping 14

1.3.4 Positional Cloning 14

1.3.5 Discovery of the Duchenne Muscular Dystrophy (DMD) disease gene 15

1.3.6 Positional candidate cloning 15

1.3.7 Current techniques 16

1.4 Next generation sequencing 16

1.4.1 Next generation technologies 16

1.4.1.1 Roche/454 FLX Pyrosequencer 17

1.4.1.2 Applied Biosystems SOLiD™ Sequencer 17

1.4.1.3 Illumina Genome Analyser 18

1.4.1.4 Life Technologies Ion Torrent PGM 18

1.4.2 Whole genome sequencing by next generation technologies 18

1.4.3 Capture and next generation sequencing 19

1.4.3.1 Array based capture 20

1.4.3.2 Solution based capture 21

1.4.3.3 Exome capture 22

1.4.4 Ethical issues of genetic screening and next generation sequencing 22

1.5 Discussion: Current and future identification of novel disease genes in Western Australia 24

Chapter 2 27

Identification of mutations in *KBTBD13* as the cause of chromosome 15 core-rod disease (NEM6) and functional analysis investigation of KBTBD13 protein

2.1 Summary 28

2.2 Introduction 28

2.2.1 Chromosome 15 core-rod disease 28

2.2.2 Family Information 32

2.2.3 Genetics of chromosome 15 core-rod disease 36

2.2.4 KBTBD13 36

2.2.5 Kelch proteins 37

2.2.6 Aims 38

2.3 Materials and methods 39

2.3.1 Screening of *KBTBD13* 39

2.3.2 Microsatellite haplotype analysis of three disease families 40

2.3.3 Screening of normal controls 40

2.3.4 Creation of plasmid constructs for transfection of mammalian cell culture 41

2.3.5 Creation of plasmid constructs for expression of KBTBD13 protein in *E. coli* 41

2.3.6 Functional studies 42

2.3.6.1 KBTBD13 protein expression in *E. coli* 42

2.3.6.2 Expression time course 43

2.3.6.3 Western blotting protocol 43

2.3.6.4 Optimising solubility of KBTBD13 protein 44

2.3.7 Transfection studies of KBTBD13 45

2.3.8 Polyclonal antibody design for KBTBD13 46

2.3.9 Assessment of antibody 46

2.3.9.1 Sample preparation 46

2.3.9.2 Western blot assessment of antibody 46

2.4 Results 47

2.4.1 Screening of *KBTBD13* in the four known disease families 47

2.4.2 Microsatellite haplotype analysis 49

2.4.3 Screening of *KBTBD13* in a series of core-rod myopathy probands 49

2.4.4 Screening of a panel of 'normal' individuals for the presence of the three known *KBTBD13* mutants 51

2.4.5 Expression studies of KBTBD13 protein in E. *coli* 51

2.4.6 Localisation of KBTBD13 within HEK-293FT cells 53

2.4.7 Localisation of KBTBD13 within C2C12 myoblasts 53

2.4.8 Polyclonal antibody production 57

2.5 Discussion 59

2.5.1 Mutation prevalence 59

2.5.2 Relationship to known kelch protein mutations and possible pathobiology 59

Chapter 3 62

Construction and implementation of next generation sequencing bioinformatic pipelines

3.1 Summary 63

3.2 Introduction 63

3.2.1 Next generation sequencing and bioinformatics 63

3.3 Elements of next generation sequencing bioinformatics pipelines 64

3.3.1 DNA quality 64

3.3.2 Library preparation 65

3.3.3 Target enrichment 67

3.3.4 Next generation sequencing 67

3.3.5 Raw data outputs 70

3.3.6 Mapping and assembly 70

3.3.7 Variant calling 71

3.3.8 Variant annotation 72

3.3.9 Variant filtering 72

3.3.10 Variant analysis 72

3.3.11 Variant interpretation 73

3.4.1 Complete Pipelines developed 73

3.4.2 Initial and intermediate pipelines 74

3.4.3 ANNOVAR-based pipelines 78

3.4.4 Additional analyses 80

3.4.4.1 Copy number variant calling 80

3.4.4.2 Synonymous variant analysis 80

3.4.4.3 Common variant filtering 81

3.4.4.4 Combing linkage or exclusion analysis with NGS data to increase the ability to identify causative variants. 81

3.4.5 Variant prioritisation 82

3.4.6 Application of the bioinformatics pipeline: Diagnostic exome sequencing example 82

3.5 Discussion 83

Chapter 4 85

Successful application of the bioinformatic pipelines to novel disease gene discovery and expanding the phenotype of known disease genes

4.1 Summary 86

4.2 Introduction 87

4.3 Materials, methods and results 88

4.3.1 Identification of *KLHL40* as a disease gene causative of foetal akinesia: (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd). 88

4.3.2 Identification of *KLHL41* as a disease gene causative of foetal akinesia (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd). 90

4.3.3 Identification of *SPEG* as a disease gene causative of centronuclear myopathy with dilated cardiomyopathy (Lead researchers Dr Gianina Ravenscroft, Prof Nigel Laing) 91

4.3.4 Expanding the phenotype of disease caused by mutations in *GBE1* (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd) 92

4.3.5 Expanding the phenotype of disease caused by mutations in *ECEL1* (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd) 93

4.3.6 Expanding the phenotype of disease caused by mutations in *CFL2* (Lead researchers Prof Nigel Laing, Mr Royston Ong, *BSc*) 94

4.4 Discussion and conclusions 95

Chapter 5 **97**

Next generation sequencing methods applied to disease gene discovery in hereditary spastic paraplegia and cerebellar ataxia

5.1 Summary 98

5.2 Introduction 98

5.2.1 Hereditary spastic paraplegia 99

5.2.2 Known pathobiology of HSPs 100

5.2.3 Inherited spinocerebellar ataxias 103

5.2.3.1 Dominantly inherited spinocerebellar ataxias 103

5.2.3.2 Polyglutamine expansion SCAs 104

5.2.3.3 Autosomal recessive spinocerebellar ataxias 104

5.2.4 Aims 107

5.3 Materials and methods: Families investigated 107

5.3.1 Family 1: HSP 107

5.3.2 Family 2: Autosomal dominant SCA 107

5.3.3 Family 3: Recessive SCA 108

5.4 Materials and Methods 110

5.4.1 Linkage analysis for families 1 and 3 110

5.4.2 Linkage analysis for family 2 110

5.4.3 Exome capture and next generation sequencing 111

5.4.4 Bioinformatic analysis of exome sequencing data 111

5.4.5 Selection of candidate variants in family 1 111

5.4.6 Screening of the SCA10 gene by Genescan in family 2 112

5.4.7 Selection of candidate variants in family 3 112

5.4.8 Sanger validation of variants in family 3 112

5.5 Family 1 results 113

5.5.1 Linkage exclusion analysis by SNPs in family 1 113

5.5.2 Exclusion of known HSP disease genes by exome sequencing 113

5.5.3 Analysis of variants identified by exome sequencing in family 1 113

5.5.4 Sanger validation of variants 114

5.6 Family 2 results 118

5.6.1 Linkage exclusion analysis by SNPs in family 2 118

5.6.2 Screening of the SCA10 gene by Genescan 118

5.6.3 Analysis of variants identified by exome sequencing in family 2 118

5.7 Family 3 results 118

5.7.1 Linkage exclusion analysis by SNPs in family 3 – autosomal recessive SCA 118

5.7.2 Exclusion of known ataxia disease genes by exome sequencing 119

5.7.3 Selection and verification of candidate variants from exome sequencing 119

5.7.4 TMEM33 119

5.8 Discussion 120

5.8.1 Family 1: X-linked HSP 120

5.8.2 Family 2: Autosomal dominant SCA 120

5.8.3 Family 3: Autosomal recessive SCA 121

5.8.4 Applications of NGS – difficulties in identifying novel disease genes 122

Chapter 6 **123**

Development and validation of a high-throughput neurogenetic sub-exomic sequencing capture panel for research and diagnostic applications

6.1 Summary 124

6.2 Introduction 124

6.2.1 Neurogenetic disorders and challenges to diagnosis 124

6.2.2 Aims 127

6.3 Materials and Methods 127

6.3.1 Capture panel design 128

6.3.2 Capture and Sequencing on the Life Technologies Proton 129

6.3.3 Variant calling and annotation 130

6.3.4 Coverage analysis 131

6.3.5 Copy number variation analysis 131

6.3.6 Sanger validation 132

6.4 Results 133

6.4.1 Sequencing statistics 133

6.4.2 Positive control hit rate 134

6.4.3 Prospective sample hit rate 134

6.4.3.1 Definitive mutations 134

6.4.4 Coverage statistics 137

6.4.5 Additional sequencing findings 137

6.4.6 Additional diagnostic findings 138

6.5 Discussion 140

6.5.1 Benefits of using a gene panel containing a large number of known disease genes 140

6.5.2 Cost effectiveness 141

6.5.3 Reasons for the success rate in identifying mutations of only around 34% in diagnostic samples 142

6.5.4 Future trends in gene panel sequencing molecular diagnostics for genetic neurological disorders 144

Chapter 7 147

Titin mutations identified as a cause of recessive minicore disease

7.1 Summary 148

7.2 Introduction 148

7.2.1 Titin 149

7.2.2 Known titin diseases 150

7.2.3 Titin skeletal muscle diseases 152

7.2.4 Aims 154

7.3 Materials and Methods 155

7.3.1 Family 1 clinical features 155

7.3.2 Family 2 clinical features 155

7.3.3 Family 3 clinical features 157

7.3.4 Additional Families 157

7.3.5 Family 1 linkage exclusion 158

7.3.6 Exome sequencing 158

7.3.7 Neurogenetic sub-exomic sequencing array 158

7.3.8 Bioinformatic processing 158

7.3.9 Variant confirmation 159

7.3.10 Molecular modelling of titin mutants 159

7.4 Results 160

7.4.1 Linkage exclusion 160

7.4.2 Family 1 160

7.4.3 Family 2 161

7.4.4 Family 3 161

7.4.5 Additional families 161

7.4.6 Variant distribution 162

7.4.7 Modelling the possible structural effect of the disease-associated variants. 162

7.4.7.1 Family 1 162

7.4.7.2 Family 2 166

7.4.7.3 Family 3 166

7.4.7.4 Mutation summary in the first three families 166

7.4.8 Additional families 167

7.4.9 Genotype-phenotype correlations 167

7.5 Discussion 170

Chapter 8 174

Final Discussion

8.1 Final Discussion 175

8.1.1 Shift from classical disease gene discovery to NGS-based gene discovery 175

8.1.2 Bioinformatic tools and databases applied to NGS 177

8.1.3 Success rates of exome sequencing-based NGS 179

8.1.4 Diagnostic applications 180

8.1.5 Emerging technologies 181

8.1.6 Final comments 183

Appendices 204

Appendix A 205

Appendix B 209

LIST OF FIGURES AND TABLES

Table 1.1: Disease genes discovered by the Molecular Genetics Laboratory, prior to 2010. 25

Figure 2.1: A: Modified Gomori-Trichrome stain of a muscle biopsy from an affected member of the Australian-Dutch family showing nemaline rods and large accumulations of rods. **B:** Electron micrograph of affected muscle from a member of the Australian-Dutch family showing a core-like area consisting of accumulations of nemaline rods. 30

Figure 2.2: A,B: Oxidative stains of patient muscle biopsy (mitochondria staining dark blue) showing the cores inherent to the disease. Figures taken from Pauw-Gommans et al., 2006.[89] 31

Figure 2.3: Pedigree of the Dutch family 33

Figure 2.4: Pedigree of the Australian-Dutch family 34

Figure 2.5: Pedigree of the Western Australian-Belgian family 35

Figure 2.6: Pedigree of the Spanish family 35

Figure 2.7: SSCP run on 37.5:1 polyacrylamide gel of 12 members of the original Dutch family. Lanes labelled A denote an affected member, U denote an unaffected member. Lanes marked –ve and +ve are negative and positive controls, respectively. 48

Fig 2.8: Microsatellite haplotypes of three affected individuals from the Australian-Dutch, Australian-Belgian and Dutch families. This figure was used in Sambuughin *et al.,* 2010.[96] 50

Figure 2.9: Chromatogram of the mutation found in an isolated Victorian proband. The mutated nucleotide is marked with an arrow. 50

Figure 2.10: Multiple sequence alignment of wild type and mutated amino acid residue in the isolated proband against 4 species. 50

Figure 2.11: Western blot of KBTBD13 protein time-course expression in E. coli. 52

Figure 2.12: Western blot of time-course expression of NusA-tagged KBTBD13 protein. 52

Figure 2.13: A HEK-293FT cell transfected with wild-type KBTBD13/DsRed. 54

Figure 2.14: C2C12 myoblasts transfected with wild-type KBTBD13/DsRed. 55

Figure 2.15: Differentiated C2C12 myotubes transfected with wild-type KBTBD13/DsRed. 56

Figure 2.16: Western blots of two antibodies produced against KBTBD13. 58

Figure 3.1: Flow diagram of an NGS pipeline from DNA sample input to useable variant data out. 69

Figure 3.2: A) Process of SOLiD sequencing ligation chemistry, with B) diagram of how 2-base encoding covers each base twice after 5 rounds of sequencing. 69

Figure 3.3: Flow diagram of the first bioinformatic pipeline constructed, initially used to analyse data from a single exome from an Illlumina sequencer. 76

Figure 3.4: Graphical overview of the intermediate pipeline produced to analyse both Illumina and SOLiD exome data. 77

Figure 3.5: Flow diagram of the latest version (V4) of the ANNOVAR variant annotation and filtering pipeline. 79

Table 5.1: Table of known HSP loci/genes with chromosomal location and mode of inheritance. 101

Table 5.2: The following table lists all currently known autosomal dominant SCA loci and disease genes, where known. 105

Table 5.3: Table of recessive ataxias with cerebellar involvement, data taken from the Neuromuscular Disorders Gene Table 2014 freeze (www.musclegenetable.fr) 106

Figure 5.1: Pedigree of family 1, consisting of three generations, with two affected individuals in the second generation. 109

Figure 5.2: Pedigree of family 2, consisting of four extant generations, with six affected individuals. 109

Figure 5.3: Pedigree of family 3, consisting of two generations, with two affected sibs in the second generation. 109

Figure 5.4: Pedigree of family 1 after clinical re-examination. 116

Figure 5.5: Chromatograms of family members showing hemizygous c.651G>C mutations in affected males and heterozygous variants in affected females. 117

Table 6.1: Table of positive control hit-rate. 136

Table 6.2: Summary table of molecular diagnoses achieved per disease class, grouped by clinical diagnosis. 136

Figure 6.1: IGV screenshot of the single coding exon of FKRP 139

Figure 6.2: IGV screenshot of the repetitive exon cluster in the NEB gene 139

Figure 7.1: Schematic of TTN protein domains, compared with an electron micrograph of muscle sarcomeres. 151

Figure 7.2: Pedigree of Family 1. 156

Figure 7.3: A, B: Oxidative stains of patient skeletal muscle showing cores and minicores (lighter areas). **C:** Fibre-typing (ATPase, pH 4) showing a predominance of type-1 fibres (stained fibres). **D:** H&E stain showing darker areas corresponding to cores (arrow) **E, F:** Electron micrographs of patient muscle showing ultrastructural detail of minicore regions in longitudinal (E) and transverse (F) sections. 156

Table 7.1: Summary table of the mutations found within the 11 studied families. 163

Figure 7.4: Schematic diagram of the TTN mutations found in 11 separate families, showing mutation name, position and predicted effect in the regions of the protein. 164

Figure 7.5: Cartoon of two titin Z1Z2 domains in antiparallel (teal, green) bound to a telethonin protein (purple), modelling the arrangement of the protein domains in the Z-disk. 165

Table 7.2: Table of genotype-phenotype correlations between the 11 patients. 169

LIST OF ABBREVIATIONS

ACP33	Maspardin, acidic cluster protein, 33-kD	LGMD2J	Limb-girdle muscular dystrophy 2J
ACTA1	Skeletal muscle alpha-actin	LMNA	lamin A/C
AD	Autosomal dominant	LOD	Logarithm of Odds
AFG3L2	ATPase family gene 3-like 2	LSBFG	Lotterywest State Biomedical Facility Genomics
AGRF	Australian Genome Research Facility	LYST	Lysosomal trafficking regulator
AMPD2	Adenosine monophosphate deaminase-2	MAG	Myelin associated glycoprotein
AP4B1	Adaptor-related protein complex 4, beta-1 subunit	MARS	Methionyl-tRNA synthetase
AP4E1	Adaptor-related protein complex 4, epsilon-1 subunit	MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)
AP4M1	Adaptor-related protein complex 4, mu-1 subunit	MRI	Magnetic resonance imaging
AP4S1	Adaptor-related protein complex 4, sigma-1 subunit	mRNA	Messenger RNA
APTX	Aprataxin	MT-ATP6	Mitochondrially encoded ATP synthase 6
AR	Autosomal recessive	MTM1	Myotubularin
ARL6IP1	ADP-ribosylation-like factor 6-interacting protein 1	MTNR1A	Melatonin receptor 1A
ARSI	Arylsulfatase I	MYH7	Myosin heavy chain 7, cardiac muscle, beta
ATL1	Atlastin	NEB	Nebulin
ATM	Ataxia-telangiectasia mutated gene	NGS	Next generation sequencing
ATN1	Atrophin 1	NIPA1	Non imprinted in Prader-Willi/Angelman syndrome 1
ATXN1	Ataxin 1	NOP56	NOP56 ribonucleoprotein
ATXN10	Ataxin 10	NSES	Neurogenetic sub-exomic sequencing
ATXN2	Ataxin 2	NT5C2	5'-nucleotidase, cytosolic II
ATXN3	Ataxin 3	PNPLA6	Patatin-like phospholipase domain containing 6
ATXN7	Ataxin 7	OMIM	Online Mendelian Inheritance in Man
ATXN8	Ataxin 8	OPA3	Optic atrophy 3
ATXN8OS	Ataxin 8 oppposite strand	PBS	Phosphate buffered saline
B3GALNT2	Beta-1,3-N-acetylgalactosaminyltransferase 2	PCR	Polymerase chain reaction
B4GALNT1	Beta-1,3-N-acetylgalactosaminyltransferase 1	PDHB	Pyruvate dehydrogenase (lipoamide) beta
bam	Binary alignment/map	PDYN	Prodynorphin
BEAN1	Brain expressed, associated with NEDD4, 1	PEO1	Progressive External Ophthalmoplegia 1
TK2	Thymidine kinase 2, mitochondrial	PEVK	Proline, glutamate, valine, and lysine
BICD2	Bicaudal D homolog 2 (*Drosophila*)	PEX7	Peroxisomal biogenesis factor 7
BMD	Becker muscular dystrophy	PGAP1	Post-GPI attachment to proteins 1
BSCL2	Berardinelli-Seip congenital lipodystrophy 2 (seipin)	PGN	Paraplegin

C12orf65	Chromosome 12 open reading frame 65	PHYH	Phytanoyl-CoA 2-hydroxylase
C19orf12	Chromosome 9 open reading frame 12	PLP1	Proteolipid protein 1
ADCK3	aarF domain containing kinase 3	PMP22	Peripheral myelin protein 22
CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	POLG	Polymerase (DNA directed), gamma
CAPN3	Calpain 3	PPP2R2B	Protein phosphatase 2, regulatory subunit B, beta
CCT5	Chaperonin containing TCP1, subunit 5 (epsilon)	PRKCG	Protein kinase C, gamma
CFL2	Cofilin-2	PVDF	Polyvinylidene fluoride
CHRNA1	Cholinergic receptor, nicotinic, alpha 1 (muscle)	RAB3GAP2	RAB3 GTPase activating protein subunit 2
CHRND	Cholinergic receptor, nicotinic, delta (muscle)	REEP1	Receptor accessory protein 1
CHRNG	Cholinergic receptor, nicotinic, gamma (muscle)	REEP2	Receptor accessory protein 2
CNV	Copy number variant	RER1	Retention in endoplasmic reticulum 1
CT	Computerised tomography	RFLP	Restriction fragment length polymorphism
CVS	Chorionic villus sample	RTN2	Reticulon 2
CYP2U1	Cytochrome P450, family 2, subfamily U, polypeptide 1	RYR1	Ryanodine receptor
CYP7B1	Cytochrome P450, family 7, subfamily B, polypeptide 1	SACS	Sacsin molecular chaperone
DDHD1	DDHD domain containing 1	sam	Sequence alignment/map
DDHD2	DDHD domain containing 2	SCA	Spinocerebellar ataxia
DMD	Dystrophin gene	SDS	Sodium dodecylsufate
DMD	Duchenne muscular dystrophy	SEPN1	Selenoprotein N1
DMEM	Dulbecco's Modified Eagle Medium	SETX	Senataxin
DNA	Deoxyribose nucleic acid	SIL1	SIL1 nucleotide exchange factor
DOK7	Docking protein 7	SilVA	Silent Variant Analysis using random forests
DTT	Dithiothreitol	SLC16A2	Solute carrier family 16, member 2
ECEL1	Endothelin converting enzyme like 1	SLC33A1	Solute carrier family 33 (acetyl-CoA transporter), member 1
EDTA	Ethylenediaminetetraacetic acid	SMN1	Survival of motor neuron 1
EGFP	Enhanced green fluorescent protein	SMN2	Survival of motor neuron 2
EM	Electron microscopy	SMRT	Single-molecule real-time
ENTPD1	Ectonucleoside triphosphate diphosphohydrolase 1	SNP	Single nucleotide polymorphism
ERLIN1	ER lipid raft associated 1	SOD1	Superoxide dismutase 1 gene
ERLIN2	ER lipid raft associated 2	SPAST	Spastin
FA2H	Fatty acid 2-hydroxylase	SPEG	Striated preferential expressed gene
FGF14	Fibroblast growth factor 14	SPG20	Spastic paraplegia 20
FIG4	Polyphosphoinositide phosphatase	SPTBN2	Spectrin, beta, non-erythrocytic 2
FKRP	Fukutin related protein	SSCP	Single-strand conformation polymorphism

FLRT1	Fibronectin leucine rich transmembrane protein 1	ssDNA	Single-stranded DNA
FnIII	Fibronectin type III	SURF4	Surfeit 4
FXN	Frataxin	SYNE1	Spectrin repeat containing, nuclear envelope 1
GAD1	Glutamate decarboxylase 1 (brain, 67kDa)	TBP	TATA box binding protein
GATK	Genome Analysis Tool Kit	TCAP	Telethonin
GBA2	Glucosidase, beta (bile acid) 2	TDP1	Tyrosyl-DNA phosphodiesterase 1
GBE1	Glycogen branching enzyme 1	TECPR2	Tectonin beta-propeller repeat containing 2
GJC2	Gap junction protein, gamma 2, 47kDa	TFG	TRK-fused gene
GPM6B	Glycoprotein M6B	TGM6	Transglutaminase 6
HMERF	Hereditary myopathy with early respiratory failure	TMEM33	Transmembrane protein 33
HSP	Hereditary spastic paraplegia	TTBK2	Tau tubulin kinase 2
HSPD1	Heat shock 60kDa protein 1 (chaperonin)	TTN	Titin
Ig	Immunoglobulin-like	TTPA	Tocopherol (alpha) transfer protein
IGV	Integrative Genomics Viewer	TVC	Torrent Variant Caller
ITPR1	Inositol 1,4,5-trisphosphate receptor, type 1	UCSC	University of California Santa Cruz
KBTBD13	Kelch repeat and BTB (POZ) domain containing 13	USP8	Ubiquitin specific peptidase 8
KCNC3	Potassium voltage-gated channel, Shaw-related subfamily, member 3	vcf	Variant call format
KCND3	Potassium voltage-gated channel, Shal-related subfamily, member 3	VNTR	Variable number tandem repeats
KIF1A	Kinesin family member 1A	VPS37A	Vacuolar protein sorting 37 homolog A (S. cerevisiae)
KIF1C	Kinesin family member 1C	WDR48	WD repeat domain 48
KIF5A	Kinesin family member 5A	ZFHX4	Zinc finger homeobox 4
KLHL40	kelch-like family member 40	ZFR	Zinc finger RNA binding protein
KLHL41	Kelch-like homologue 41	ZFYVE26	Zinc finger, FYVE domain containing 26
KLHL9	Kelch-like family member 9	ZFYVE27	Zinc finger, FYVE domain containing 27
L1CAM	L1 cell adhesion molecule		

AIMS

This thesis had four main aims:

1) To use the classical gene discovery methods of positional candidate cloning to identify the genetic basis of the dominantly inherited core-rod myopathy previously linked to chromosome 15q.

2) To research, develop and implement a next generation sequencing bioinformatics pipeline.

3) To use the bioinformatics pipeline in next generation sequencing-based gene discovery methods to investigate the genetic basis of a number of neurogenetic diseases, for which the causative genes were not yet known.

4) To research, develop and implement next generation sequencing based diagnostics for neurogenetic diseases and genetic cardiomyopathies.

Chapter 1

Introduction: genetic diseases, methods of disease gene discovery

1.1 Aims

This thesis had four main aims:

1) To use the classical gene discovery methods of positional candidate cloning to identify the genetic basis of the dominantly inherited core-rod myopathy previously linked to chromosome 15q.

2) To research, develop and implement a next generation sequencing bioinformatics pipeline.

3) To use the bioinformatics pipeline in next generation sequencing-based gene discovery methods to investigate the genetic basis of a number of neurogenetic diseases, for which the causative genes were not yet known.

4) To research, develop and implement next generation sequencing based diagnostics for neurogenetic diseases and genetic cardiomyopathies.

1.2 Genetic disease

Genetic disease is defined as any illness caused by abnormalities in nuclear or mitochondrial DNA. These mutations may be inherited or they may be acquired de novo.[1]

The total number of genetic disorders is vast, with the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim) recording a total of 22,693 individual entries as of 11th December 2014, spanning all the autosomes, the sex chromosomes and the mitochondrial genome. Of these entries, 4,305 have a phenotypic description with a known molecular basis i.e. the disease gene is known, an additional 88 describe a variant in association with a known phenotype, 1,850 have a phenotype with a suspected Mendelian bias and 1,677 have a phenotype description or locus, but no molecular diagnosis *i.e.,* no causative gene has been identified.[2] This clearly

demonstrates that while there are many known disease genes, there are still many more to find.

Potentially, most genes in the human genome are candidates to cause disease. MacArthur *et al.* describe that each individual harbours approximately 20 genes that we humans can apparently do without; genes which have been shown to harbour homozygous or compound heterozygous loss-of-function mutations, yet the person appears unaffected by this lack of gene function. The presumption is that there are a large set of genes that are redundant, with the function of the protein products of these genes not essential for "normal" function, perhaps because of duplicated copies of the gene, or the ability of genes with similar function to compensate for the loss.[3] Hence the importance of furthering knowledge about genetic diseases, finding causative disease genes and ascertaining the mechanisms by which mutations in these genes cause disease.

1.2.1 The importance of finding a causative disease gene

The identification and characterisation of the causative disease gene is a crucial event for any genetic disease.

In order to locate a human disease gene, two things are required. First, a map of the human genome is needed with the positions and features of known genes and variants. One such map can be found at the UCSC Genome Browser.[4] The second requirement is a cohort of patients and/or families who have the disease to study. Australia is a good place to find mutations causing dominant disease, as large families that have married out are relatively common, allowing for highly informative pedigrees. A good example of a large family with a dominant disease is one affected by familial amyotrophic lateral

sclerosis, which at the time was one of the largest ever published, and led to the identification of mutations in the *SOD1* gene as the first known cause of familial amyotrophic lateral sclerosis.[5] Similarly the single large family with dominantly inherited childhood onset nemaline myopathy through which mutated slow alpha-tropomyosin was identified as the first known gene for nemaline myopathy[6] and the family that led to the identification of mutations in the myosin heavy chain 7 gene (*MYH7*) as the cause of Laing distal myopathy.[7; 8] Recessive disease does not have a high incidence in the Australian population, as there is a relatively low incidence of consanguineous marriages, although this is changing with recent migration patterns.(Laing, NG, Personal communication)

Prior to the identification of the disease gene often all that is known is the disease phenotype. Genetic counselling may even be difficult as the exact mode of inheritance in a family may be unclear. For example, what might appear to be a recessive disease, with an affected child of unaffected parents, may in fact be the result of a *de novo* dominant mutation in the child. This was the case when mutations were first identified in the skeletal muscle alpha-actin gene (*ACTA1*)[9]. Thus, the mode of inheritance may only be known for sure once the disease gene and precise mutation in any given patient is identified, and that information may alter the genetic counselling as to the risks for subsequent pregnancies.

The identification of a disease gene:

1) Allows accurate molecular genetic diagnosis of the disease.
2) May allow prognosis.
3) Allows prenatal or presymptomatic diagnosis where appropriate.

4) Allows investigation of the pathobiology of the disease based on knowledge of the disease gene and mutated protein.
5) Increases understanding of the biology of the affected tissue, for example muscle in a muscle disease, especially if the protein is previously unknown.
6) Allows the rational development of therapies based on understanding of the precise pathological mechanism resulting from the mutation.
7) Further understanding of tissue biology from characterising a novel disease gene may lead to therapeutic interventions for other, related diseases.

1.2.1.1 Accurate Diagnosis

Identification of a disease gene initially leads to a genetic diagnosis of disease in the patient(s) or family(ies) it was first discovered in, enabling all family members to be screened for the mutation. The discovery then enables the population of other patients and families with a similar disease phenotype to be screened for mutations in the new disease gene. This can lead to a large number of genetic diagnoses where none was previously possible. An accurate molecular diagnosis can avoid the diagnostic odyssey, and give psychosocial benefits to the patients and their families. Clinicians may also request a molecular test first, rather than subjecting the patient to an invasive procedure such as a muscle biopsy.[10-12]

1.2.1.2 Prognosis

Following accurate diagnosis and genetic testing, a prognosis for the disease may be given, based on known genotype/phenotype correlation data. For some diseases, there may be a typical clinical progression, allowing the clinician to give the patient a prognosis and enabling the patient to plan ahead and arrange care or assistance if needed. In the case of the neurogenetic diseases, often there is significant variability in

disease progression and phenotype, even between patients with the same mutation. For example, the exact same deletion in dystrophin (*DMD*) may result in either severe Duchenne muscular dystrophy (DMD) or mild Becker muscular dystrophy (BMD), one of the best known examples of this is the relatively common exon 3-7 deletion which may result in either DMD or BMD in individual patients depending on how the patient splices round the mutation.[13; 14] A disease where clinical severity can be predicted with better accuracy is spinal muscular atrophy caused by homozygous inactivation of the survival of motor neuron 1 (*SMN1*) gene. The disease severity is known to be influenced by the copy number of *SMN2*, an almost identical gene, and is inversely correlated with *SMN2* copy number and the level of full-length SMN protein produced by *SMN2*.[15-17]

On the other hand, for diseases without a known genetic cause, it would be even more difficult to give an accurate prognosis, or even a diagnosis, leading to much uncertainty as to the future prospects of the patient.

1.2.1.3 Prenatal, preimplantation and presymptomatic diagnosis

Prenatal diagnosis involves testing for genetic diseases before birth, to determine whether the unborn child has inherited the known family genetic disease mutation(s). The most accurate genetic prenatal diagnosis is only possible when the disease-causing mutation is identified in a patient or family. Prenatal genetic diagnosis usually involves testing a chorionic villus sample (CVS) for the presence of the family mutation[18; 19] or pre-implantation genetic diagnosis.[20]

Initial prenatal diagnosis for some diseases may be done by ultrasound and blood testing of the mother for specific features that may indicate disease e.g. absence of foetal

movement.[21] These initial safe tests can then be used to help determine whether more invasive prenatal genetic tests (CVS, amniocentesis) are needed[22; 23].

Preimplantation genetic diagnosis is an alternative to prenatal diagnosis. In preimplantation diagnosis, the at-risk couple undergo in vitro fertilisation, and then, before an embryo is implanted, several cells are removed and analysed to look for the family mutation.[24; 25]

Presymptomatic diagnosis for genetic diseases is performed when a proband within a family is identified, and a causative mutation is found. Family members can then be screened for the presence of the mutation prior to the onset of symptoms. This allows them to make necessary life decisions with more certain knowledge of whether or not they will develop the disease. The first genetic disease for which presymptomatic diagnosis was established was Huntington disease. The Huntington's protocol of pre-test psychological evaluation, testing two samples and post-test counselling was established to minimise the psychological impact on the patients and their families.[26]

Presymptomatic diagnosis is frequently performed for late onset neurodegenerative disorders such as Huntington disease[27] or familial motor neuron disease caused by mutations in the superoxide dismutase 1 gene (*SOD1*).[5] However, the variable penetrance of disease alleles in these cases, and other late-onset, variably penetrant diseases complicates the counselling in presymptomatic diagnosis. In the case of *SOD1* motor neuron disease one recommendation was that because of the difficulty in counselling due to incomplete penetrance, pre-symptomatic *SOD1* genetic testing should only be performed after detailed genealogical investigations have been performed.[28] In the case of Huntington's, a CAG trinucleotide repeat expansion to

between 36-39 repeats results in the reduced penetrance of the Huntington phenotype.[29] A more recent study recommended that individuals with repeat expansions in the intermediate (27-35) range should be considered at risk of developing Huntington's disease.[30]

1.2.1.4 Investigation of the pathobiology of genetic disease

The discovery of a disease gene allows the investigation of the pathobiology of the mutant protein, and determination of how that particular mutation may cause the disease. The mechanisms may be more easily ascertained when the disease gene has a product of known structure and function, as with some ACTA1 mutants. Although some mutations have been functionally characterised, the pathobiology of others remain unknown.[9] There are also cases where both structure and gene function is known, for example *MYH7*, where the pathobiology of mutants is still unknown.[31] For an older example, mutations in *SOD1* were described as causative of familial amyotrophic lateral sclerosis in 1993,[32] yet even in 2014 researchers are still attempting to elucidate the full mechanism of disease pathobiology.[33; 34] If a novel disease gene codes for a previously unidentified or uncharacterised protein, the difficulty of discovering the pathobiology behind the disease will be likely significantly increased.

1.2.1.5 Increasing understanding of tissue biology

Implication of a gene in disease can also significantly expand understanding of the biology of the tissue or tissues affected by the disease. This has occurred most notably when the identified disease gene codes for a previously unknown protein. An early example of this was the dystrophin protein, when it was first implicated in the severe X–linked childhood disease Duchenne muscular dystrophy (DMD). Hoffman *et al.,*[35] sequenced approximately 25% of the total 14-kb DMD transcript from human foetal

skeletal muscle and mouse adult heart, and found that not only were the transcripts 90% identical, but suggested that the protein could serve a structural role in muscle tissue, which we now know is the case.

Another area where knowledge of muscle biology is still increasing even now is in how the architecture of the sarcomere is arranged and regulated. Mutations in the nebulin gene were first identified in 1999;[36] at the time there was an incomplete understanding of its function in the sarcomere, though it was known to be a component of the sarcomeric thin filament. Since then, two different models of the function of nebulin in regulating the sarcomere have been proposed. The first paper proposed that nebulin protein functions as a molecular ruler and was the major determinant of thin filament length;[37] The second paper proposed that nebulin was part of a two-segment model, where one thin filament segment (bound to nebulin) was static in length, and the second variable length segment was governed by microregulation of actin dynamics.[38]

1.2.1.6 Rational development of therapies

Therapies for genetic diseases can be difficult to implement as the correction of the root cause, the genetic mutation itself is difficult, although the disease itself can be treated; for example, a heart transplant in someone affected with a cardiomyopathy.

The discovery of a disease gene may also lead to the development of new therapies based on understanding the precise genetic pathology of the disease. One example of the development of therapy based on knowledge of the disease-causing mutations is anti-sense oligonucleotide-induced exon skipping in the dystrophin gene. These oligonucleotides mitigate the effects of mutations by causing skipping over the exon(s)

that are affected and the production of partially functional dystrophin protein.[39] Anti-sense oligonucleotide-induced exon skipping is now in clinical trials[40; 41]

On the other hand, an exon skipping technique would be useless if applied to myopathies caused by actin mutations, as actin is a very small protein and all six protein coding exons are critical for its function.[42] However, as another example, myopathies caused by skeletal muscle actin null mutants have been corrected in mouse models by the expression of cardiac actin in skeletal muscle, allowing a related form of the protein to compensate for the loss of skeletal muscle actin[43].

Viral-mediated gene therapy has had a long history, first being used to treat severe combined immunodeficiency to mixed success – four patients out of ten in one clinical trial developed leukaemia from the retroviral vector inserting near an oncogene.[44] A more recent success more relevant to muscle disease is the successful application of viral-mediated gene therapy to a catecholaminergic polymorphic ventricular tachycardia mouse model, a single injection to the heart providing a curative effect for one year after injection.[45]

It remains to be said that development of therapies and treatments for genetic diseases must be tailored to the specific gene, mutation type and the delivery method of the treatment.

1.2.1.7 Development of other therapeutics

Identifying the pathobiology of a disease may lead to candidate therapies for other disorders. For example, if a novel gene mutation causing a pain insensitivity disorder is

found, a drug could be developed to mimic the pain insensitivity effects of the mutation for a novel or more effective method of anaesthesia[46].

1.3 Identification of disease genes – A short historical perspective

1.3.1 Gene discovery before linkage maps and the Human Genome Project

The concept of genetic linkage (the tendency for certain loci or alleles to be inherited together) dates back to the early 20th century, when, in fruit flies, Thomas Morgan first described a sex-linked trait[47]. Morgan also proposed that the frequency of recombination between two loci (then called 'crossing over') could be an indication of the distance separating two genes on the chromosome[48].

The concept of linkage in regards to the mapping of a disease gene to a quantitative trait locus dates back to the 1950s, with the first set of equations to detect and analyse autosomal linkage within a family being published by Morton.[49] These equations also give the LOD (Logarithm of ODds) score; which defines the probability that two loci are linked. LOD scores are used as a measure of the probability that a candidate region on the genome obtained by linkage analysis is genuine.

As a statistical measure, positive LOD scores provide evidence in favour of linkage and negative LODs give evidence against linkage. At a single point, a LOD score of 3 is the threshold for accepting linkage, corresponding to 1000:1 odds, which although seemingly strict corresponds to a Bayesian calculation of a conventional p=0.05 threshold of statistical significance.[50] Linkage can be rejected when LOD score of -2 is returned.[51] This however is for the analysis of a single point. For a genome-wide measure of significance, Lander & Schork[52] suggest a LOD score of 3.3 is the minimum threshold to indicate linkage for Mendelian characters, which corresponds to having 10

informative phase-known meiotic events, either in one family or from adding different families together.[52]

Before the advent of the Human Genome Project and in the absence of a linkage map with complete coverage of the genome, initial studies hinged upon tests of cosegregation between the disease phenotype and a trait inherited in a Mendelian manner.[49; 53] This approach was fraught with difficulty, given that the coverage of the genome by the available markers was very low.[54] It was not until the construction of linkage maps of the whole genome that the frequency of disease gene discovery started to increase.

1.3.2 Construction of human linkage maps

The first protocol for producing a 'dense' linkage map was published in 1980 by Botstein *et al.,*[55] in which they suggested that at least 150 restriction fragment length polymorphism (RFLP) sites would be needed for sufficient coverage to map the entire genome, though more information would be gained with a greater number of markers.[55] These sites would be detected by cleavage of human genetic material with restriction endonucleases, where the length of some of the fragments detected (determined by agarose gel electrophoresis) would change based on the genotypic differences between individuals that would modify the enzyme recognition site. From that digestion step, fragments encoding specific sequences could be probed by Southern blot, and associated length polymorphisms identified. More importantly, these markers could be associated with chromosomal locations, and as such, begin to map out the human genome. This method was a paradigm shift, in that, because it provided a map of the entire genome, enabled researchers to dissect where on the genome a given quantitative trait locus resided, and whether the inheritance of certain patterns of RFLP bands was

linked with a genetic disease. RFLP mapping however is very time consuming,[56] and as such was supplanted by microsatellite analysis.[57] Microsatellites, otherwise known as variable number tandem repeats (VNTRs) are short repetitive regions of variable length occurring throughout the genome. Linkage of a disease gene to microsatellites is done by comparison of the pattern of allele lengths in affected and unaffected individuals in families. Microsatellite analysis, although still used up until just before the start of my thesis work, has now given way to array-based single nucleotide polymorphism (SNP) analysis as the most common method, allowing thousands to millions of markers with defined positions across the entire genome to be analysed at once.[58; 59]

1.3.3 Exclusion mapping

Even with the first construction of the RFLP linkage map, there were still often situations where the family structure was not suitable for linkage analysis, and even with microsatellite and SNP linkage analysis today, this can still be a problem. However, if a disease has multiple candidate genes, a large number of these can be eliminated using linkage results in the process of exclusion mapping,[60-62] where all regions with a LOD score of below -2 are considered to be statistically 'excluded' from being the candidate region.[51]

1.3.4 Positional Cloning

Positional cloning is the identification, or cloning of a disease gene purely by positional bioinformatic information such as linkage data from a large family or a cohort of patients, without any information on the biochemical basis of the disease. This approach to disease gene discovery works even when there is little or no information about the biochemical basis of the disease available.[63]

1.3.5 Discovery of the Duchenne Muscular Dystrophy (DMD) disease gene

The first example of linkage mapping leading to the discovery of a disease gene was the discovery of mutations of the dystrophin gene.[64] The exact localisation of the gene was discovered using data obtained from three different sources, an example of positional cloning:

1) Several reports described females suffering from a classical DMD phenotype, who were found to have a translocation of a portion of the X chromosome to an autosome, with the breakpoint in the region of Xp21.
2) The discovery of DNA markers linked to the Duchenne gene, and then the localisation of these markers, and the gene itself using RFLP markers.
3) A boy was reported with DMD in conjunction with chronic granulomatous disease, retinitis pigmentosa, and McLeod syndrome. Upon analysis, he was found to have a small deletion in Xp21, which presumptively included the Duchenne gene.[65]

1.3.6 Positional candidate cloning

The positional candidate approach as described by Collins[66] was then not a major approach to human disease genetics, the majority of disease genes having been discovered by either positional or functional approaches. Between 1986 and 1995, 42 disease genes had been identified by the positional cloning method.[66] The positional candidate approach involves mapping a disease locus to a region of the genome by linkage analysis, and then conducting a survey of the linkage region to see if any likely candidate genes resided there. Collins[66] predicted that this would soon become the preferred method of disease gene discovery, and that prediction has held true. This approach is still currently a 'best practice' method of disease gene discovery.

1.3.7 Current techniques

At the commencement of this thesis, the 'best practice' methods for disease gene discovery were a refinement of the 'positional candidate' approach supported by Collins[66] where genome-wide linkage analysis is performed within a family or families segregating the disease under investigation in an attempt to define a candidate region where the gene of interest could be. Once this candidate region has been defined, a list of genes within the region can be produced from the draft human genome website[4] (http://genome.ucsc.edu/) and candidate genes based on known gene homologies or known pathway interactions can be selected and analysed. There are now various bioinformatic programs to aid in the selection of candidate disease genes, based on common disease phenotype, intra-cellular localisation of the protein expression profiling and protein structure, for example, Gentrepid.[67] However, this approach is likely to quickly fall by the wayside, or be used as an additional analysis instead of a primary one with the availability of next generation sequencing technologies.

1.4 Next generation sequencing

1.4.1 Next generation technologies

The availability of affordable commercial next generation sequencing technologies from 2004 onwards[68] has led to a paradigm shift in the search for novel disease genes. The most striking example is that instead of having to painstakingly gather a family or families large enough to produce significant linkage and perform exclusion analysis of known disease genes, it is possible to make a diagnosis from single small families, and isolated probands. [69; 70]

1.4.1.1 Roche/454 FLX Pyrosequencer

This system was the first of the so-called next generation sequencing platforms to become commercially available, in 2004.[71] It uses an alternative to Sanger sequencing called pyrosequencing, the principle being that the incorporation of a nucleotide by DNA polymerase results in the release of a pyrophosphate molecule, which initiates a series of downstream reactions that ultimately produce light via the enzyme luciferase. The amount of light produced is proportional to the number of nucleotides incorporated.[68]

The 454 system uses a fragmented DNA library to which adapters are ligated, from which a single-stranded DNA (ssDNA) library is formed. These ssDNA fragments are then annealed to agarose beads with surface oligonucleotides complementary to the adapters. The beads are then mixed with PCR reagents in a water-in-oil emulsion, and the mix is loaded onto a microtitre plate, a flat plate with multiple tiny wells, such that each well contains a single bead, thus creating the microreactors in which amplification occurs. Finally, the emulsion is broken and each bead is sequenced by the pyrosequencing technique in individual wells.[68]

1.4.1.2 Applied Biosystems SOLiD™ Sequencer

Like the Roche/454 system, the SOLiD platform uses a fragmented, adapter-ligated library and an emulsion PCR amplification step, but it differs in the method of sequencing. It uses DNA ligase and specific 8 base-pair long colour encoded primers as a basis for its sequencing.[72]

1.4.1.3 Illumina Genome Analyser

Like the other two systems mentioned, the Illumina system uses a fragmented, adapter-ligated library, but omits the emulsion PCR step and instead uses bridge amplification of immobilised ssDNA fragments to create clusters of fragments on the surface of the flow cell to which the fragments are bound. The Illumina system utilizes a sequencing by-synthesis approach in which all four nucleotides are added simultaneously to the flow cell, along with DNA polymerase, for incorporation into the cluster fragments. Each of the nucleotides is blocked at the 3' and carries a specific fluorescent tag, and an imaging step follows each incorporation step. After imaging, the fluorophore and the block are removed, and the cycle continues.[73]

1.4.1.4 Life Technologies Ion Torrent PGM

The Life Technologies Ion Torrent sequencer is one of a new generation of non-optical NGS machines. While still operating on a sequencing-by-synthesis approach, the method of detection is via semiconductor-based hydrogen ion detection; in essence, each well functions as a micro pH meter. Individual nucleotides are flowed across the surface of the detector chip, where bead-bound single stranded DNA is immobilised in the detection wells. The magnitude of the electric charge detected can then be processed to detect the number of bases added during each nucleotide flow.[74]

1.4.2 Whole genome sequencing by next generation technologies

The first published human genome to be sequenced by next generation methods was published in 2008 by Wheeler *et al.,*[75] using the FLX 454 sequencing technology. This was the sequence of a single individual and within that genome a large number of variations were found. In total, 3,322,093 SNPs were found, 2,715,296 were known changes, and 606,797 were novel. Of the 3.3 million SNPs, 10,654 were non-

synonymous, and 1,743 of those were novel. Also identified were a large number of insertions, deletions and copy number variant (CNV) regions.[75]

Ultimately, routine sequencing of whole genomes for study will become common as computing power and the cost of sequencing go down in price, due to economies of scale and increasing technological power.[76] Currently, however, it is uneconomical to sequence whole genomes routinely because the time taken and the cost of materials and burden upon the informatics infrastructure in most research labs are too great.[76] There have been examples however of sequencing of a human genome to find disease genes,[77] and though it is still relatively expensive, with the decreasing cost of next generation sequencing it will become more common.

As a consequence of this, considerable effort has been devoted to the development of high-throughput capture and enrichment technologies to enable selective sequencing of regions of interest in the genome.

1.4.3 Capture and next generation sequencing

Since the commencement of this thesis, sequencing technologies have advanced to the point that sequencing of adequate numbers of whole genomes with sufficient coverage within a disease family is cost effective, the X Ten system from Illumina offering full coverage human genomes for less than $1,000 USD.[78]

However as analysis of such large sets of data is still not time-effective, there is a need for targeted capture and enrichment of sections of the human genome. The original targeted enrichment procedure was PCR, with 1 reaction per amplicon targeting one region. The logical extension of this type of targeted resequencing is multiplex PCR,

where multiple amplicons are generated per reaction, up to an entire exome with the Ion AmpliSeq Exome kit from Life Technologies. (https://www.lifetechnologies.com/order/catalog/product/4487084)

Another method for enriching a large number of targets are molecular inversion probes, where single-stranded DNA sequences are annealed to a target with target-specific flanking sequences, and the DNA is circularised with a ligase prior to PCR amplification and sequencing. This platform offers cost-effectiveness in high-throughput scenarios of up to ~10 000 exons, along with ease of use and high sensitivity, specificity and reproducibility.[79]

Currently, the most high-throughput methods of targeted capture and enrichment are the hybrid capture family of techniques, consisting of either on-array or in-solution capture methods.

1.4.3.1 Array based capture

Array based capture uses a shotgun fragment library made from the DNA sample of interest, hybridised to a probe set immobilised on a chip. After hybridisation, non-specific hybrids are washed off to leave the regions of interest hybridised to the arrayed capture probes.

The first on-array capture chip was made by Roche NimbleGen[80] and their collaborators. The first generation of their chip was able to capture 4-5Mb of sequence. Their second generation HD2 chip was able to capture up to 34Mb. Initially these chips were designed for use with the Roche 454 sequencers, but protocols were later produced to optimise the chips for use with other next generation sequencing technologies.[76]

The array capture chips do have drawbacks. The chief of these is the high cost of the hardware needed to handle the chips, and the need for a relatively large amount of DNA, between 10 and 15µg, irrespective of the size of the region to be captured.[76; 80] Also there is a physical limit to the number of arrays that can be hybridised by a single person per day, even working with samples in parallel. This time consumed does not include the necessary analysis and assembly steps needed to transform the raw read data into a format where variants can be identified.

1.4.3.2 Solution based capture

To overcome the limits of array-based capture, both Agilent and NimbleGen have developed methods of solution-based capture. It uses essentially the same principle as array-based capture, where specific probes are designed to target regions of interest. The solution-based capture method however uses an excess of probe compared to target, as opposed to an excess of target used in the on-chip method of capture and the probes are bound to magnetic beads, instead of being bound to an immobile substrate.[81] The solution capture method is easily scalable, with reactions able to be run, for example, in 96-well plates.

Both methods are currently in use, however, solution-based capture is supplanting array-based capture, as it addresses many of the shortcomings of array-based capture methods: chiefly, the high cost and DNA requirements, and the lack of scalability of array-based capture.[76]

1.4.3.3 Exome capture

Exome capture is an application of the two hybrid capture techniques, where all the coding exons within the genome (the exome) are targeted, along with some of the flanking intronic sequence.[76] The human exome accounts for a little more than 1 percent of the whole genome. However the majority of pathogenic mutations are thought to reside in exons or immediate intronic flanking sequence, so exome capture and sequence will be an effective method in most cases.[82] Therefore, sequencing only the exome is relatively efficient in terms of cost and analysis time compared to an entire genome, and results in substantially less data to analyse.

It should be noted that exome capture and sequencing will not detect deep intronic mutations that may introduce splice sites as those regions are not targeted by the capture method. Large repeat expansions will also not be directly detected due to the short read lengths of most reaction chemistry. Direct detection would require a read length that spanned the expansion. Newer methods allow indirect inference of repeat expansions through computational methods, but for whole genome data only.[83]

1.4.4 Ethical issues of genetic screening and next generation sequencing

With the commercial availability of next generation sequencing and current technological progress,[78] it will soon be possible for an individual to have their entire genome screened for possible risk alleles and deleterious mutants by a providing company, with possible analysis and interpretation of the most critical results by a qualified genetic counsellor.

The situation is different when examining genomes and exomes sequenced as a diagnostic test to interrogate a limited set of genes, compared to a genomic test

performed on a research basis to identify a novel disease gene. When a genomic screen is performed in a clinical setting, a list of criteria and guidelines for reporting are available from the American College of Medical Genetics and Genomics, alleviating some of the ethical burden.[12]

In a recent (2014) paper surveying the attitudes of genetic professionals towards the return of incidental results from exome and whole-genome sequencing found that 85% of respondents thought that incidental genomic results should be offered to adult patients, and 74% thought that results should be offered to the parents of a child with a medical condition. Sixty-two percent thought that incidental results about adult-onset conditions and carrier status should be offered to the parents of a child with a medical condition, and about half thought that offered results should not be limited to those deemed clinically actionable. An overwhelming majority of 81% thought that return of results should be guided by the patient's preferences.[11]

For genomic tests performed on a research basis, reporting all such variants to the presiding clinician would require that each variant identified in the research laboratory would have to be verified independently in an accredited diagnostic laboratory before information could be passed back to the patient or research study participant. The study participant would then have to be counselled about the significance of each of the variants identified to either themselves or their offspring. Importantly, research laboratories are bound by the ethics clearances that are given or imposed upon them by their local ethics committee. In the case of the Centre for Medical Research Molecular Genetics Laboratory, the ethics clearance given mandated that we were not to feed back any incidental findings.

Finally, there is also the issue of data availability; should the participant's genome be publicly released for research purposes? The data obtained may be of importance to other research labs around the world, but it would be impossible to obtain informed consent for each and every secondary application of the research data obtained.

All these issues will need to be addressed in the coming years as next generation sequencing techniques become more and more ubiquitous.

1.5 Discussion: Current and future identification of novel disease genes in Western Australia

Until 2010, in Western Australia, disease gene discovery has been successfully achieved with the classical techniques of the positional candidate approach and subsequent candidate gene analysis. These techniques are sorely limited in cases and families where there is little information available to narrow down a candidate region or candidate disease genes. Next generation technologies therefore offer researchers a much more powerful toolset for the discovery of human disease genes, as the quantity of data able to be produced is so much higher than traditional methods, and the need for large families and significant linkage data are reduced.

Before the start of my PhD work, the classical methods of positional cloning and candidate gene cloning had been used successfully in over a dozen different human disease gene identification projects by the Centre for Medical Research Molecular Genetics Laboratory headed by Winthrop Professor Nigel Laing. The table below shows the disease genes identified by the Laboratory up to 2010.

Table 1.1: Disease genes discovered by the Molecular Genetics Laboratory, prior to 2010.

Familial motor neuron disease	*SOD1*	1994	Suthers *et al.,* 1994[84]
Dominant nemaline myopathy	*TPM3*	1995	Laing *et al.,* 1995[6]
Recessive nemaline myopathy	*NEB*	1999	Pelin *et al.,* 1999[36]
Actin myopathy, AD/AR nemaline, intranuclear rod Myopathy	*ACTA1*	1999	Nowak *et al.,* 1999[9]
Core-rod disease	*RYR1*	2000	Scacheri *et al.,* 2000[85]
Craniometaphyseal dysplasia	*ANKH*	2001	Nürnberg *et al.,* 2001[86]
Congenital fibre type disproportion	*ACTA1*	2004	Laing *et al.,* 2004[87]
Early onset distal myopathy	*MYH7*	2004	Meredith *et al.,* 2004[8]
Nemaline Myopathy	*CFL2*	2007	Agrawal *et al.,* 2007[88]
Cap disease	*TPM2*	2007	Lehtokari *et al.,* 2007[89]
Hereditary Spastic Paraplegia	*CYP7B1*	2008	Tsaousidou *et al.,* 2008[90]

The Laboratory has access to large cohorts of nemaline myopathy patients, as well as Charcot-Marie-Tooth disease patients, to name the two largest. Over the more than 25 years the Laboratory has been active over 20,000 DNA samples have been collected and are available for study. The Laboratory maintains a close collaborative relationship with the PathWest Department of Diagnostic Genomics, Western Australia, where a large number of research samples originate from.

I aim to leverage these resources, using the classical gene discovery methods of positional candidate cloning to identify the genetic basis of the dominantly inherited core-rod myopathy previously linked to chromosome 15q. I aim to use next generation technologies to aid in the discovery of disease genes in families which cannot be dealt with using the classical techniques within the Laboratory's cohorts, either in that there is not significant linkage within the family, or there are simply too many candidate genes to screen within the linkage region. In order to do that, I will need to research, develop and implement a next generation sequencing bioinformatics pipeline, and use that bioinformatics pipeline in next generation sequencing-based gene discovery methods. As a final implementation of next generation sequencing, I aim to research, develop and implement a next generation sequencing based diagnostic screen for neurogenetic diseases and genetic cardiomyopathies.

Chapter 2

Identification of mutations in *KBTBD13* as the cause of chromosome 15 core-rod disease (NEM6) and functional analysis investigation of KBTBD13 protein

2.1 Summary

Three mutations causative of chromosome 15 core-rod disease were identified in the gene kelch repeat and BTB (POZ) domain containing 13 (*KBTBD13*) in four families using the traditional method of positional candidate gene screening. Subsequently, a third mutation was discovered in an isolated proband from Victoria Australia upon screening *KBTBD13* in 14 further core-rod disease probands using Sanger sequencing. KBTBD13 protein is expressed in *E. coli* and two mammalian cell lines, and an attempt is made to create a polyclonal antibody to the protein. This was the first characterisation and description of the disease gene *KBTBD13* and associated mutations that cause NEM6. This is also the first characterisation of KBTBD13 protein and associated mutations that cause NEM6.

2.2 Introduction

2.2.1 Chromosome 15 core-rod disease

Chromosome 15 core-rod disease (NEM6, OMIM #609273) is a rare form of dominantly inherited nemaline myopathy first described in a Dutch family by Gommans *et al.* (2002).[91] The locus for the disease gene was then linked to chromosome 15q21-23 in 2003 using the original Dutch family, and an Australian-Dutch family.[92]

Upon clinical examination, NEM6 displays a mild to moderate phenotype with muscle histopathology displaying both nemaline rods and central cores upon staining and electron microscopy (Figures 2.1, 2.2). The onset of disease varies from late adolescence to early adulthood and is slowly progressive, initially manifesting as a difficulty in climbing stairs or getting out of a low chair. In general, the muscle weakness seems to begin proximally, but may extend to a diffuse display of weakness over time. No cardiac features have been seen in any of the studied families.[92; 93] The

Dutch and Australian-Dutch families show prominent weakness in the neck flexors and proximal limb muscles, displaying a limb-girdle pattern of muscle weakness. Recent clinical data also indicated phenotypic similarity between these two published families and the Spanish[94] and Western Australian-Belgian families (P. Lamont, unpublished data). Involvement of facial, respiratory, ankle dorsiflexion or cardiac muscles is not present in any of the four families, (P. Lamont, unpublished data).[91; 92; 94] Nerve conduction studies in the Dutch family displayed normal results; however, needle electromyography found non-specific myopathic features that were more pronounced in the elderly.[91] Muscle computerised tomography (CT) scans in the Dutch family, found that fatty infiltration of the muscle was only slight or absent in younger members of the family, whereas definite fatty infiltration of the muscle was present in the older patients.[91] In the Australian-Dutch family, MRI scans of the proband of the family indicated symmetrical muscle wasting and fatty replacement.

Unique to the NEM6 phenotype is a form of muscle 'slowness', with the patients having no prompt motor responses to sudden or unexpected events[91; 92; 94] This is an interesting feature that could not be detected on regular neurological examination or upon eletromyographic examination,[93] however, the patients upon examination display a lower contraction speed and rate of relaxation, as well as a lower torque generation compared to healthy controls.[93]

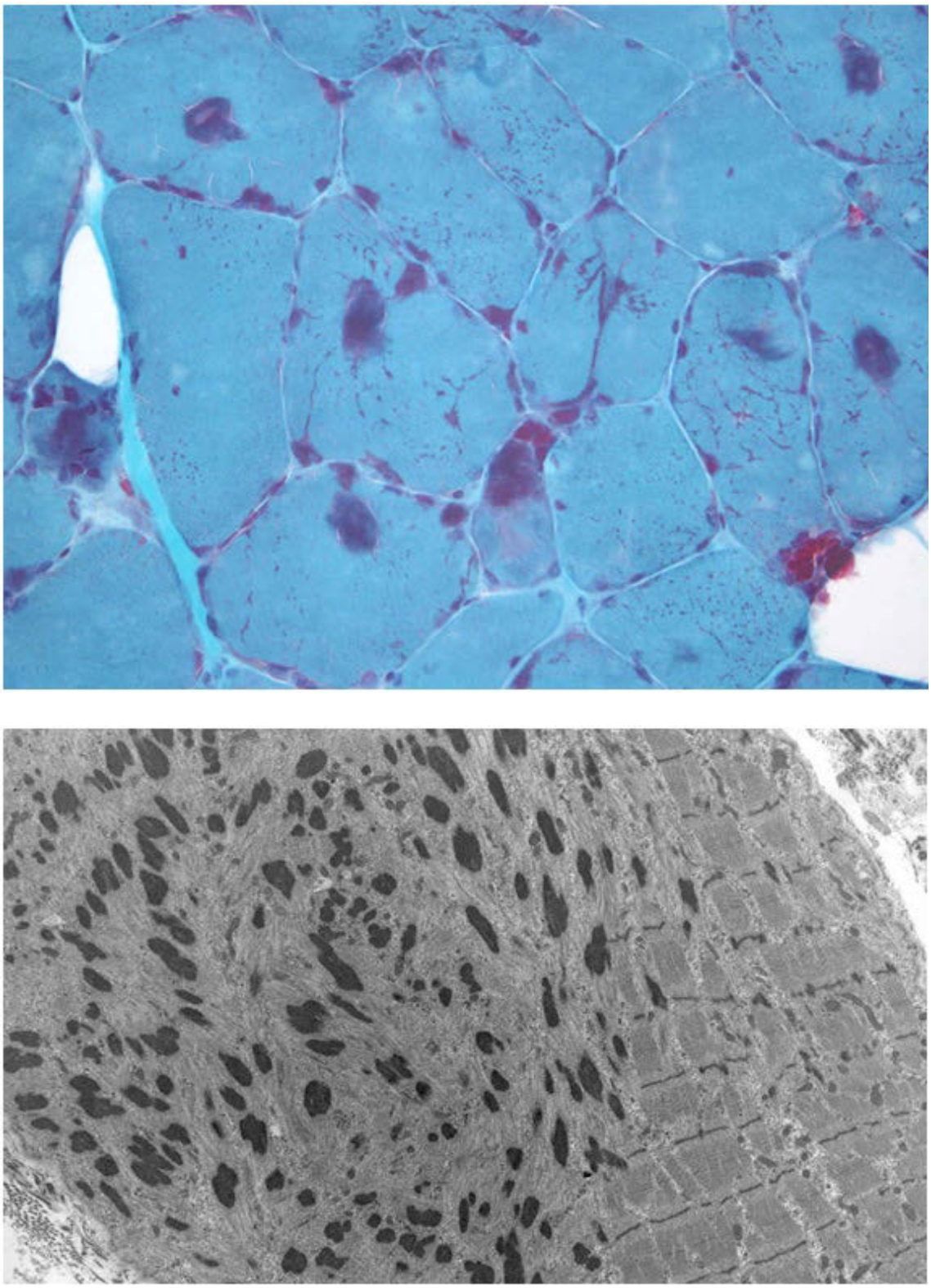


Figure 2.1: A: Modified Gomori-Trichrome stain of a muscle biopsy from an affected member of the Australian-Dutch family showing nemaline rods and large accumulations of rods. **B:** Electron micrograph of affected muscle from a member of the Australian-Dutch family showing a core-like area consisting of accumulations of nemaline rods.

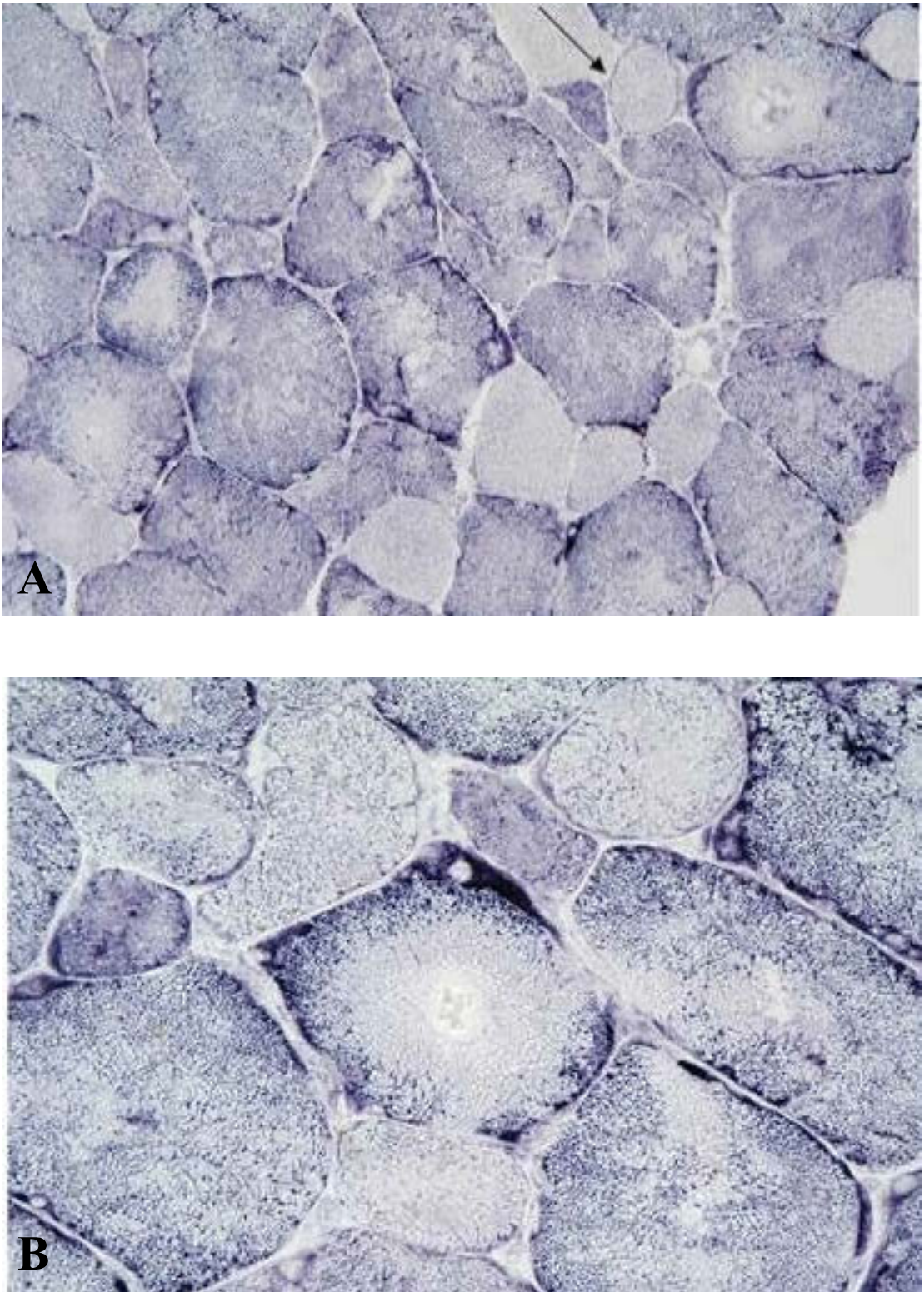


Figure 2.2: A,B: Oxidative stains of patient muscle biopsy (mitochondria staining dark blue) showing the cores inherent to the disease. Figures taken from Olive *et al.,* 2010.[94]

2.2.2 Family Information

Four families are available for this study, the original Dutch and Australian-Dutch families, described by Gommans *et al*. in 2002 and 2003,[91; 92] respectively, along with recently identified Spanish and Western Australian-Belgian families.

The Spanish family is a three-generational family consisting of 12 studied members, four of whom are affected. The Western Australian-Belgian family is a 4 generational family consisting of 7 studied members, 4 of which are confirmed to be affected. Both the Spanish family and the Western Australian-Belgian family were published in Sambuughin *et al.,* 2010.[95] See Figure 2.3 for the pedigree of the Dutch family, Figure 2.4 for the pedigree of the Australian-Dutch family, Figure 2.5 for the pedigree of the Western Australian-Belgian family and Figure 2.6 for the pedigree of the Spanish family.

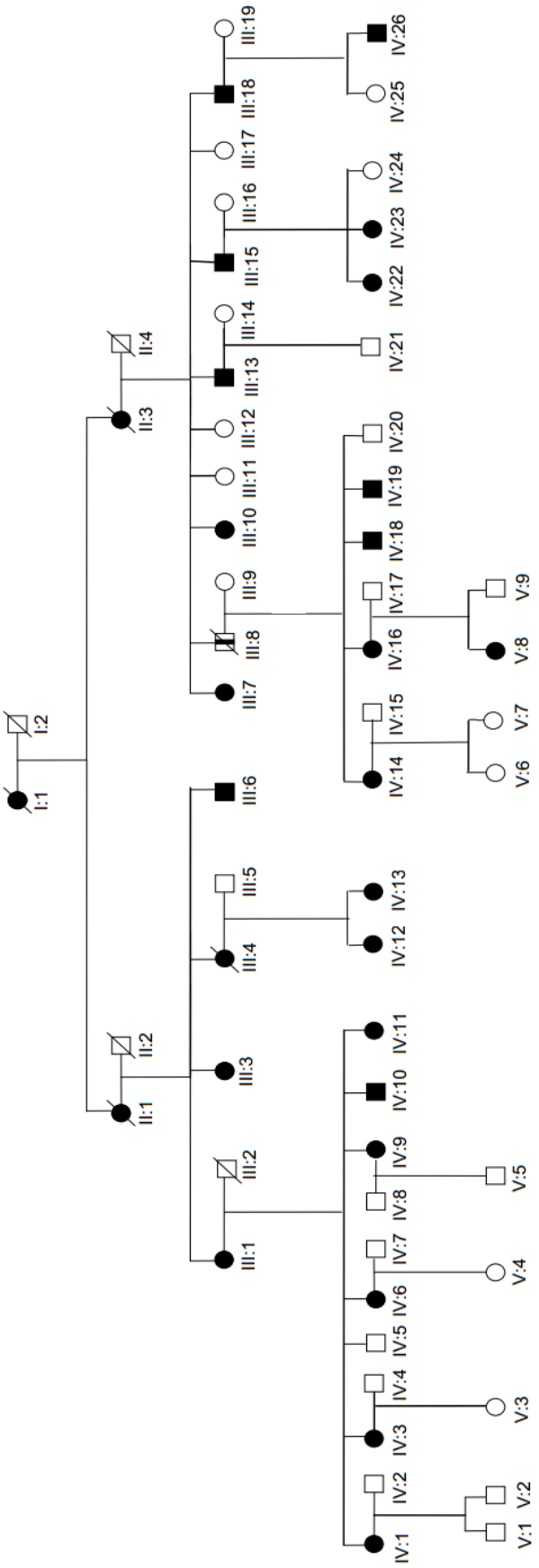


Figure 2.3: Pedigree of the Dutch family

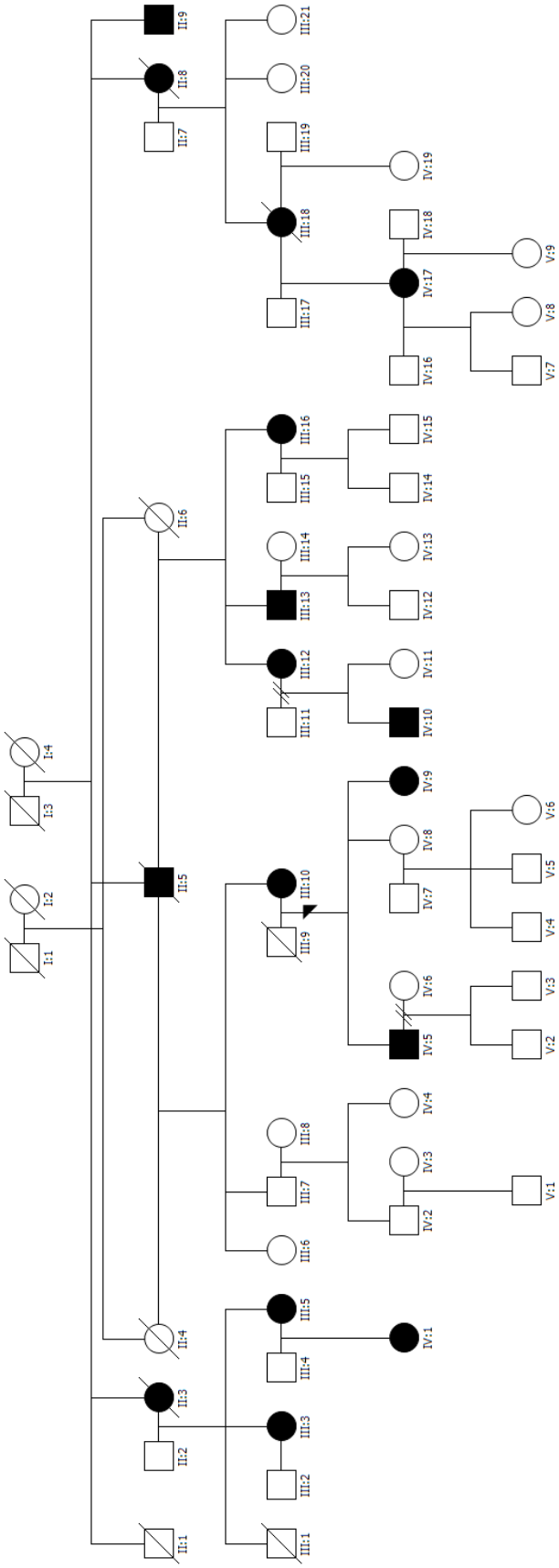


Figure 2.4: Pedigree of the Australian-Dutch family, with proband indicated by an arrow.

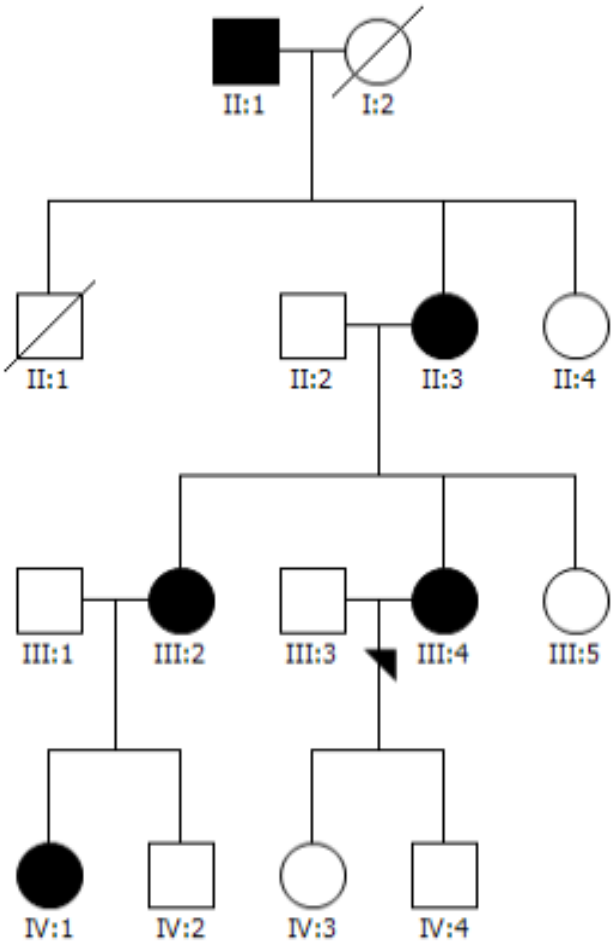


Figure 2.5: Pedigree of the Western Australian-Belgian family, with proband indicated by an arrow.

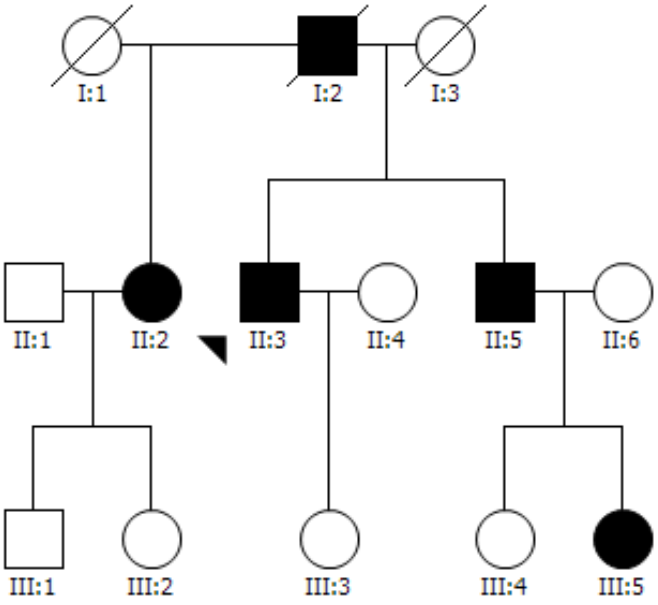


Figure 2.6: Pedigree of the Spanish family, with proband indicated by an arrow.

2.2.3 Genetics of chromosome 15 core-rod disease

As previously stated, the locus for the disease gene was linked to chromosome 15q21-23 in 2003 using the original Dutch family, and an Australian-Dutch family, between the microsatellite markers D15S1033 and D15S151 in an affecteds-only analysis, and between the markers D15S155 and D15S125 when both affected and unaffected family members were analysed.[92]

In collaboration with the Goldfarb group (National Institutes of Health, Bethesda, Maryland, U.S.A.), 28 candidate genes within the linkage region were screened, of which I screened eight. In early 2010, the Goldfarb group suggested that they had discovered the causative gene as one of the candidate genes they were screening as their part of the collaboration. The Goldfarb group had discovered two different variants in a putative gene labelled *LOC390594*, now designated *KBTBD13*: one in the Spanish family, c.1170G>C, causing a change p.Lys390Asn, and another variant c.1222C>T in the Australian-Dutch family from the DNAs that we supplied them with, causing a change p.Arg408Cys.

2.2.4 KBTBD13

KBTBD13 is a single-exon gene encoding a BTB-Kelch protein of unknown function and structure. Protein databases and computational modelling predicted that it consists of a BTB/POZ (Bric-a-brac tramtrack Broad-complex/Pox virus and Zinc finger) domain and a β-propeller formed of 5 - 6 repeated kelch folds.

2.2.5 Kelch proteins

Kelch proteins are a superfamily of proteins with diverse roles in cells. The definitive feature of kelch proteins is the presence of a beta-propeller structure formed from up to seven kelch repeats. The kelch repeat displays very little sequence homology between different kelch-repeat containing proteins, with as little as 11% homology between different kelch proteins in *Drosophila*.[96; 97] Of particular interest given the identification of *KBTBD13* as a disease gene for NEM6 is the biological function of BTB-kelch proteins. The BTB domain is the Broad-Complex, Tramtrack and Bric a brac domain; also known as a Poxvirus and Zinc-finger (POZ) domain.

A 2010 publication describes a mutation in a related BTB-Kelch protein, KLHL9, as the cause of a form of early-onset distal myopathy.[98] However, the mutation described is within the BTB domain of the protein, as opposed to the mutations found in KBTBD13 which were all found within the kelch β-propeller. Cirak *et al.*[98] determined that the mutation in the BTB domain of the protein was interfering with the binding of the KLHL9 protein to Cullin 3, preventing the formation of the E3 ubiquitin ligase complex. A similar mechanism of pathology may be in action with NEM6, where instead of the mutation interfering with the formation of the E3 ubiquitin ligase complex, mutations in the β-propeller kelch domain may be interfering with, or altering the substrate-binding capacity of KBTBD13.

2.2.6 Aims

This chapter aims to:

(1) Screen *KBTBD13* in the Dutch and Australian-Belgian families to determine what mutation they have and whether it segregates perfectly with the disease phenotype.

(2) Screen *KBTBD13* in core-rod disease probands without a known genetic cause from the cohort at the RPH Neurogenetic Unit.

(3) Express fluorescently-tagged KBTBD13 in mammalian cell culture to determine a possible localisation for the protein.

(4) Design an antibody specific to KBTBD13 protein in order to localise the expression of KBTBD13 protein in cultured myoblasts and human muscle biopsies.

(5) Express and purify soluble KBTBD13 protein in order to study its binding partners by immunoprecipitation assays, once a suitable antibody has been produced.

2.3 Materials and methods

2.3.1 Screening of *KBTBD13*

KBTBD13 was screened by PCR amplification and sequencing. The gene was amplified in 3 fragments of approximately 600bp in length, with approximately 50bp of overlap between each fragment. The Qiagen Q-solution PCR mix and the Slowdown 70 thermocycling protocol were used to amplify all three fragments. Primer sequences can be found in Table 2.1 of Appendix A, PCR mixes and thermocycling conditions can be found in Appendix A. Affected individuals from the Australian-Dutch, Australian-Belgian, Dutch and Spanish families were screened by Sanger sequencing for the two mutations discovered by the Americans to check for segregation of the mutant with the disease phenotype.

51 members of the Dutch family for which DNAs were available were screened by SSCP to test for complete segregation. Separate primers were designed to amplify a fragment that spanned a region of 278bp that would contain both the Spanish and Dutch mutation sites. Primer sequences are in Table 2.2 of Appendix A.

3µl of each amplicon was then mixed with 9µl of formamide loading buffer and denatured of 95°C for >5 minutes, then snap-chilled on wet ice for >5 minutes. The denatured products were then run on a 12% 37.5:1 polyacrylamide gel for 1 hour at 400V and subsequently silver stained using standard protocols.

Fourteen core-rod probands of unknown genetic cause from the cohort at the Royal Perth Hospital Neurogenetics Laboratory were also screened via PCR and sequencing as per the protocol earlier in this section.

2.3.2 Microsatellite haplotype analysis of three disease families

Padma Sivadorai of the Pathwest Genetic Diagnostics lab, Department of Health Western Australia performed Microsatellite haplotype analysis on subsets of the Dutch, Australian-Dutch and Australian-Belgian families using the Genescan (Applied Biosystems) allelotyping technique.[99] The markers interrogated were D15S993, D15S1018, D15S1009, D15S108, D15S107 and D15S1020.

2.3.3 Screening of normal controls

A set of 50 normal controls available from WAIMR was screened via SSCP for the Spanish and Australian-Dutch/Australian-Belgian/Dutch mutations. The primers and SSCP protocol used to screen the Dutch family are in Table 2.2, Appendix A, and Section 2.3.1.

A set of 149 Australasian normal controls was screened by restriction enzyme digest for the NM_001101362:c.742C>A mutation found in an isolated Australasian proband. Primers were designed to encompass the mutation, which would introduce an *Alu*I (New England Biolabs, (NEB)) restriction site in addition to the single site already present in the amplicon, creating fragments of 16 and 140bp for the normal samples and fragments of 16, 65, 75 and 140bp for the mutant upon digestion. Primer sequences can be found in Table 2.3, Appendix A.

The Goldfarb group screened 53 additional normal controls for the presence of the mutation.

2.3.4 Creation of plasmid constructs for transfection of mammalian cell culture

The full-length mutant and wild-type *KBTBD13* gene was amplified from genomic DNA using KOD polymerase (Novagen) using the primers at Appendix A Table 2.4a and the KOD polymerase mix and thermocycling protocol in Appendix A. The *KBTBD13* propeller domain sequence was amplified using the primers at Appendix A Table 2.4b and the KOD polymerase mix and thermocycling protocol in Appendix A. These amplicons were then gel-purified and subsequently cloned into the Zero Blunt® TOPO® vector (Invitrogen) as per the manufacturer's instructions. Transformed cultures were plated out onto LB agar plates containing 50µg/ml kanamycin and incubated at 37°C overnight. Colonies were then picked and grown in 5ml of LB broth containing 50µg/ml kanamycin in a 37°C shaker at 220RPM overnight. Plasmid DNA was harvested using the Purelink plasmid miniprep kit (Invitrogen) and the insert was excised with *Kpn*I (NEB) and *Xho*I (NEB) restriction enzymes.

The vector dsRED2-N1 was digested with *Kpn*I and *Xho*I restriction enzymes and the *KBTBD13* insert was ligated in using T4 DNA ligase (NEB). Transformation, selection and harvesting were all carried out as per previously stated in this section.

Sequence verification of the insert and confirmation that the *KBTBD13* gene was correctly in-frame with the fluorescent tag was carried out using the primers found in Appendix A, Table 2.4).

2.3.5 Creation of plasmid constructs for expression of KBTBD13 protein in *E. coli*

Amplification and sub-cloning of KBTBD13 into Zero Blunt® TOPO® was carried out using the same procedure as in section 2.3.4, with the primers used found in Appendix

A table 2.5. For cloning into pETM-11, the insert was excised with *Kpn*I (NEB) and *Nco*I (NEB) restriction enzymes from the construct.

For cloning into pET-44a, the insert was amplified out of the construct produced in section 2.3.4 using KOD polymerase and the conditions as in section 2.3.4 with the primers found in Appendix A Table 2.6, subcloned into TOPO and excised with *Kpn*I (NEB) and *BamH*I (NEB) restriction enzymes.

The vector pET-44a (Novagen) was cut with *BamH*I (NEB) and *Kpn*I (NEB) restriction enzymes. The *KBTBD13* insert was ligated in using T4 DNA ligase (NEB) according to manufacturer's protocol. Transformation, selection and harvesting were all carried out as per previously noted in section 2.3.4. Verification of correct ligation the insert and confirmation that the *KBTBD13* gene was wild-type and correctly in-frame with the N-terminal NusA tag was carried out by restriction enzyme digest and sequencing of the construct using primers found in Appendix A Table 2.6.

2.3.6 Functional studies

Functional studies on KBTBD13 protein were carried out in parallel with the Goldfarb Group at the National Institutes of Health, Bethesda.

2.3.6.1 KBTBD13 protein expression in *E. coli*

An overnight culture of 3mL LB broth was inoculated from a glycerol stock of bacteria containing the desired expression construct. This was used to inoculate 50mL of LB broth and was grown in a 250mL Erlenmeyer flask at 37°C with shaking at 220RPM to the half-log phase (OD_{600} of 0.7 – 0.9). The culture was then cooled on ice for 20min. Protein expression was induced with the addition of IPTG to a final concentration of 0.5mM and the culture was then shaken (220RPM) at 37°C for 4-6 hours.

2.3.6.2 Expression time course

An expression time course followed by western blotting was undertaken to assess the solubility of the expressed pETM-11 6x-His/KBTBD13 protein. Aliquots of 1ml of expression culture were taken at induction (t=0) and at 1-hour intervals following induction. The aliquots were then centrifuged at top speed in an Eppendorf 5424 centrifuge for 2 minutes and the supernatant discarded. The cell pellet was resuspended in 200µl lysis buffer (0.5M NaCl containing 1mM EDTA, 1% Triton X-100, 1mM DTT and bacterial protease inhibitor cocktail) and incubated on ice for 20-30 minutes before sonication for two minutes at 30W using a Qsonica Q55 sonicator with 3.2mm tip.

Lysates were then centrifuged at maximum speed for 10 minutes to pellet the insoluble fraction. The supernatant containing the soluble fraction was transferred into a fresh 1.5ml eppendorf tube. The insoluble pellet was suspended in 20µl urea lysis buffer (8M urea containing 2% triton X-100, 18mM DTT, 1% glycerol and bacterial protease inhibitor cocktail) and 20µl SDS gel loading buffer containing 5% β-mercaptoethanol, mixed by pipetting and vortex, and then heated to 95°C for 10 minutes to dissolve the pellet.

The antibodies used were an anti-6x histidine tag primary antibody at 1/5000 dilution, and horseradish peroxidase-tagged anti-mouse secondary as per section 2.3.6.3

2.3.6.3 Western blotting protocol

Cell lysates and protein ladder (Invitrogen) were loaded onto a 4-12% gradient denaturing polyacrylamide gel in SDS running buffer (Invitrogen) and electrophoresed

at 200V for 90 minutes. The protein was then blotted onto a PVDF membrane (Thermo Scientific) at 300mA for 120 minutes in tris-glycine buffer containing 20% methanol.

Membrane with immobilised protein was blocked in ~40mL blocking solution (150mL phosphate buffered saline (PBS) containing 0.1% TWEEN20, with 7.5g skim milk powder (Homebrand) added) for a minimum of one hour. After blocking, incubations with primary antibodies diluted in blocking solution were performed at 4°C overnight. Post-primary incubation, membranes were washed three times for a minimum of 30 minutes per wash in wash solution (PBS containing 0.1% TWEEN20).

Incubations with horseradish peroxidase-tagged secondary antibodies diluted in blocking solution 1/10000 were performed at room temperature for one hour. Post-secondary incubation, membranes were washed three times for a minimum of 30 minutes per wash in wash solution. Chemiluminescent detection was performed with a Pierce ECL Western Blotting Substrate kit according to manufacturers instructions. Membranes were imaged using X-ray film (Thermo Scientific) and films were developed using a Kodak X-ray film developer.

2.3.6.4 Optimising solubility of KBTBD13 protein

Optimising KBTBD13 protein expression for solubility was attempted by attaching a NusA solubility tag[100] to the C-terminus of the KBTBD13 protein, and expression of this construct with butanol-enriched medium as per the protocol found in Quan et al, 2011.[101]

Time-course expression was performed as per section 2.3.6.2, and soluble and insoluble fractions were examined by electrophoresis on a 4-12% gradient polyacrylamide gel and

western blotting using an anti-6x histidine tag primary antibody at 1/5000 dilution, and horseradish peroxidase-tagged anti-mouse secondary as per section 2.3.6.3.

2.3.7 Transfection studies of KBTBD13

Transfections were carried out in mammalian cell culture using C2C12 and HEK-293FT cells using Lipofectamine-2000 transfection reagent according to standard protocols. For both C2C12 and HEK-293FT cells, the KBTBD13/dsRED2 construct was transfected in triplicate, with a previously existing ACTA1/EGFP construct to act as a transfection control.

HEK-293FT cells were plated out 24 hours before transfection at approximately 40% confluence into 6-well culture plates (VWR) in transfection media (Dulbecco's Modified Eagle Medium (DMEM) (GIBCO) containing 5% foetal calf serum (FCS) and 10mM L-glutamine). After transfection, transfection media was removed and replaced with growth media (transfection media with 5mM penicillin/streptomycin (pen-strep) cocktail), and was changed every 24 hours.

C2C12 mouse myoblasts were plated out 24 hours before transfection at approximately 35% confluence in 24-well culture plates (VWR), coated with Matrigel (Corning Life Sciences) in a 1/10 dilution, applied using the manufacturer's instructions. Initial transfection was performed in transfection media, and cells were imaged 48 hours after transfection. To differentiate the transfected cells into myotubes, the media was changed to DMEM containing 10% horse serum, 10mM L-glutamine and 5mM pen-strep. All imaging was done on an Olympus IX71 fluorescence microscope.

2.3.8 Polyclonal antibody design for KBTBD13

Two polyclonal antibodies specific for KBTBD13 was produced by Kathy Davern at the (Monoclonal Antibody Facility, Harry Perkins Institute for Medical Research), specifically to the kelch domain of the protein using the epitope VRGDTVYTVNR (amino acids 398-408 of the protein). The epitope sequence is identical to the corresponding sequence in murine KBTBD13. These antibodies would then theoretically cross-react with murine KBTBD13 protein, enabling studies of KBTBD13 localisation in mouse muscle. Both antibodies were affinity purified against the epitope sequence.

The manufactured antibodies were then assessed using western blot and immunohistochemistry for specificity to KBTBD13 protein.

2.3.9 Assessment of antibody

2.3.9.1 Sample preparation

Muscle samples were prepared from frozen control muscle biopsies by cutting approximately 20 8µm slices on a cryostat, followed by solubilisation of the cut sections in 150µl of urea lysis buffer plus 50µl SDS gel loading buffer containing 5% β-mercaptoethanol. The samples were heated to 95°C for 10 minutes to dissolve the sections.

2.3.9.2 Western blot assessment of antibody

The antibody was tested for specificity by western blotting, as per the protocol in 2.3.6.3. Antibody was tested against solubilised KBTBD13 protein expressed in *E. coli*, fluorescently-tagged KBTBD13 expressed in HEK-293FT cells, HEK-293FT cell lysate and two solubilised human muscle samples, prepared as per section 2.3.9.1.

2.4 Results

2.4.1 Screening of *KBTBD13* in the four known disease families

Screening for segregation of the candidate variants with the disease phenotypes in the Spanish (NM_001101362:c.1170G>C, (p.Lys390Asn)), and Australian-Dutch (NM_001101362:c.1222C>T, (p.Arg408Cys)) families showed complete segregation of the variants with the disease in the families.

Upon screening of the Dutch and Australian-Belgian families for the presence of a *KBTBD13* variant, they were found to have the same NM_001101362:c.1222C>T (p.Arg408Cys) mutation as the Australian-Dutch family.

Segregation analysis in the Australian-Belgian family revealed complete segregation of the variant with the disease phenotype.

SSCP analysis (Figure 2.7) of all 51 members of the Dutch family for whom DNA was available found segregation of the NM_001101362:c.1222C>T (p.Arg408Cys) variant with the disease phenotype in 49 members of the family. For the two members of the family for which there was not segregation, the members underwent a second clinical examination, and member IV:26 who in figure 2.3 was designated affected was in fact not affected, and member V:7 designated unaffected was found to have developed the disease. These diagnoses are in line with the SSCP results that I obtained; therefore, after re-examination, segregation was demonstrated in all 51 members of the family.

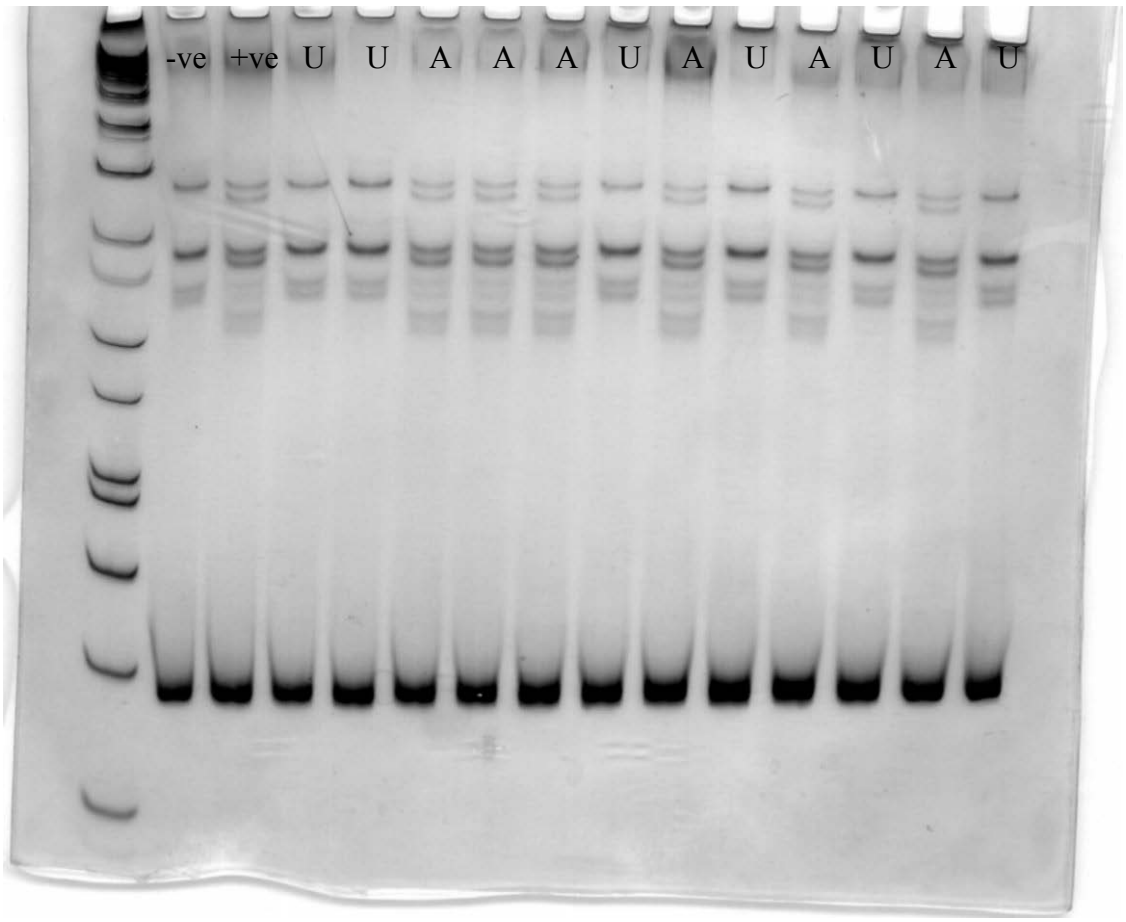


Figure 2.7: SSCP run on 37.5:1 polyacrylamide gel of 12 members of the original Dutch family. Lanes labelled A denote an affected member, U denote an unaffected member. Lanes marked –ve and +ve are negative and positive controls, respectively.

2.4.2 Microsatellite haplotype analysis

Subsequent microsatellite haplotype analysis of individuals from the Australian-Dutch, Australian-Belgian and Dutch families revealed a common founder haplotype between them, as shown in figure 2.8. A common haplotype in the Dutch and Australian-Dutch families is bounded by markers D15S993 and D15S1020. The inclusion of the Australian-Belgian family reduces the common haplotype to being bounded by markers D15S108 and D15S1020.

2.4.3 Screening of *KBTBD13* in a series of core-rod myopathy probands

Screening of 14 unrelated core-rod disease probands identified 1 mutation in an isolated proband: NM_001101362:c.742C>A, causing a change p.Arg248Ser (Figure 2.9). The proband's father is stated by family members to have had a similar disease, but no DNA was available from either parent.

The mutated residue was found to be conserved in KBTBD13 homologues across species, as seen in Figure 2.10, and protein modelling found that the mutated residue would lie within the second kelch propeller domain.

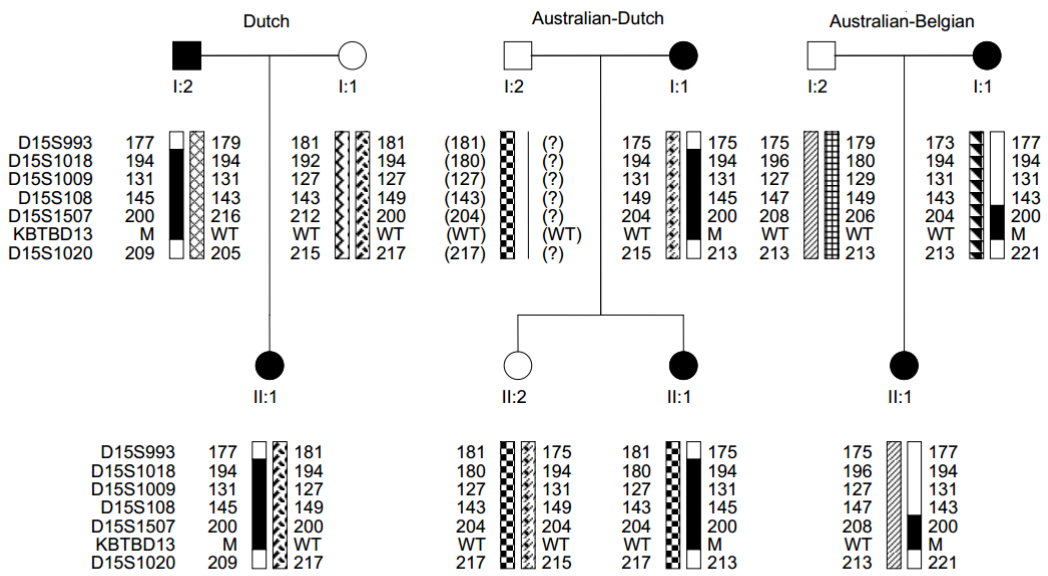


Fig 2.8: Microsatellite haplotypes of individuals from the Australian-Dutch, Australian-Belgian and Dutch families. This figure was used in Sambuughin *et al.,* 2010.[95] A large shared haplotype bounded by D15S993 and D15S1020 is present between the Dutch and Australian-Dutch families, and a smaller shared haplotype between all three families is bounded by D15S108 and D15S1020.

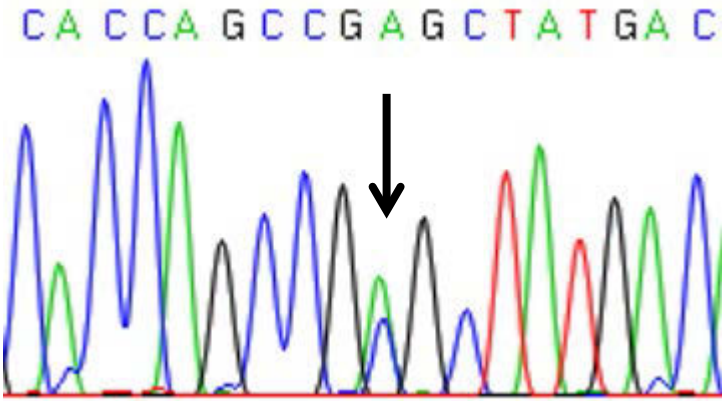


Figure 2.9: Chromatogram of the NM_001101362:c.742C>A mutation found in an isolated Victorian proband. The mutated nucleotide is marked with an arrow.

```
Human           224  VVELGFCYDPDGGTWHEFPSPHQPRYDTALAGFDGRLYAIGGEF 268
Human mutated   224  VVELGFCYDPDGGTWHEFPSPHQPSYDTALAGFDGRLYAIGGEF 268
M. mulatta      224  VVELGFCYDPDGGTWREFPSPHQPRYDTALAGFDGRLYAIGGEL 268
D. labrax       224  VVELGFCYDPEGGTWCEFPSPHQPRYDMALAGFEGRLYAIGGEF 268
A. vittatum     164  --------DADATAWSEFPSPHQLRYDVRLVGHEGYLYAIGGEY 208
D. rerio        233  I-SAAHCYNPSKNEWNQVASLNQKRSNFKLLAVSGKLYAVGGHC 277
```

Figure 2.10: Multiple sequence alignment against 4 species of KBTBD13 wild type and mutated amino acid sequence, with the mutated residue in the isolated proband highlighted in red..

2.4.4 Screening of a panel of 'normal' individuals for the presence of the three known *KBTBD13* mutants

Of the 50 Australasian controls screened by SSCP for the presence of the Spanish and Australian-Dutch/Australian-Belgian/Dutch mutations none were found to have the mutation. Combined with the 216 normal controls screened by the Goldfarb group, these data indicate that the two mutations found in these families are likely to be causative of the NEM6 phenotype.

Of the 148 normal controls I screened for the presence of the candidate variant identified in the isolated Victorian patient, none were found to have it. Combined with the 52 normal controls screened by the Goldfarb group, these data indicate that the variant found in this family is likely to be causative of the Victorian core-rod phenotype.

2.4.5 Expression studies of KBTBD13 protein in E. *coli*

In E. *coli*, KBTBD13 protein was found to be insoluble at all stages of protein production when expressed with a 6xHIS tag, as shown in the time-course in figure 2.11.

Expression of the KBTBD13/NusA construct in E. *coli* with butanol-enriched media found that the protein was insoluble at all stages of expression. (Figure 2.12)

Expression of the kelch propeller domain of KBTBD13 with an n-terminal 6xHIS tag found that the protein was insoluble at all time-points.

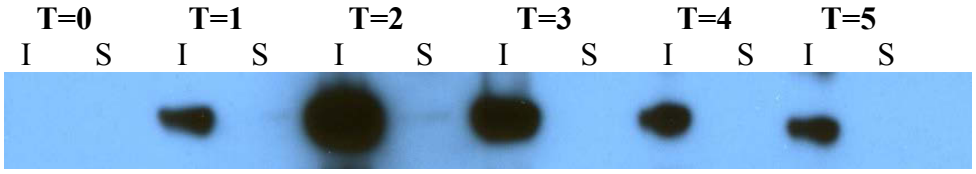


Figure 2.11: Western blot of KBTBD13 protein time-course expression in E. coli. Time points were taken at T=0, and subsequent 1 hour intervals. I stands for insoluble fraction, S stands for soluble fraction. All KBTBD13 protein product was in the insoluble fraction.

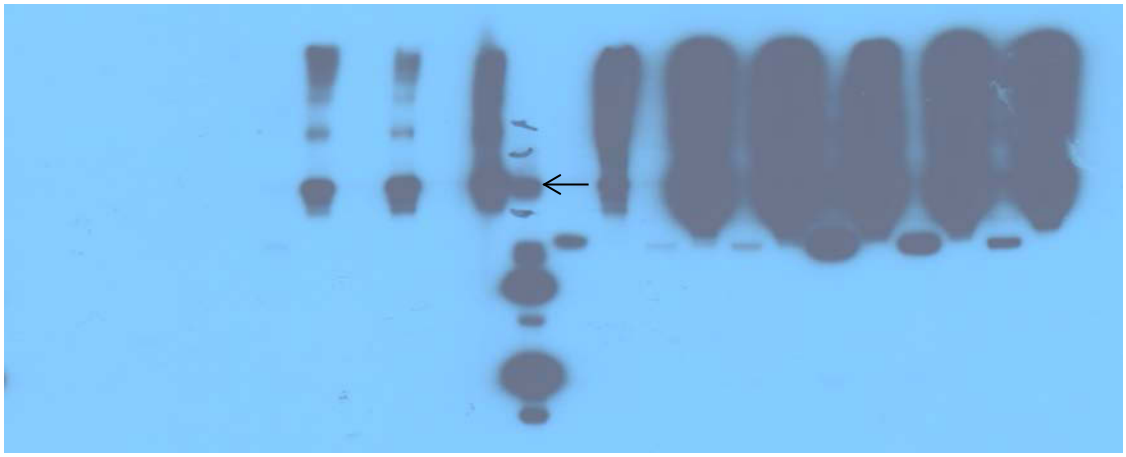


Figure 2.12: Western blot of time-course expression of NusA-tagged KBTBD13 protein. Two samples (S1/I1, S2/I2) were run on this western. Time points were taken at T=0, and subsequent 1 hour intervals. L is the ladder (Novex Sharp pre-stained, Invitrogen). The ladder was seen to react with anti-His antibodies, hence the visible band. All KBTBD13 protein product was in the insoluble fraction.

I stands for insoluble fraction, S stands for soluble fraction. The NusA tag has a MW of 54kDa. When attached to KBTBD13 would result in a total estimated protein weight of 103kDa. A black arrow indicates 110kDa on the ladder.

2.4.6 Localisation of KBTBD13 within HEK-293FT cells

Expression of dsRED2-tagged KBTBD13 protein within HEK-293FT cells showed punctate cytoplasmic expression and cytoplasmic aggregates., shown in figure 2.13. Mutant KBTBD13 protein when expressed at high levels showed the same subcellular localisation and propensity for aggregation as wild-type KBTBD13.

2.4.7 Localisation of KBTBD13 within C2C12 myoblasts

At 24 hours post-transfection, no expression of *KBTBD13*/DsRed was seen in any of the transfected wells, whereas the *ACTA1*/EGFP positive transfection control was seen to be expressing. However, upon leaving the transfection media on the cells for 48 hours, expression was seen. Examination of the myoblasts revealed that the KBTBD13/DsRed protein localised primarily around the nucleus, while also having diffuse expression in the cytoplasm. Also when expressed at high levels, large cytoplasmic inclusion bodies can be seen. These data are shown in Figure 2.14.

Upon differentiation of the transfected C2C12 cells into myotubes, it was seen that the pattern of expression of KBTBD13/DsRed protein did not change, still primarily localising around the nucleus while being expressed in lower levels in the cytoplasm. Inclusion bodies are still seen when expression is at high levels. This can be seen in Figure 2.15.

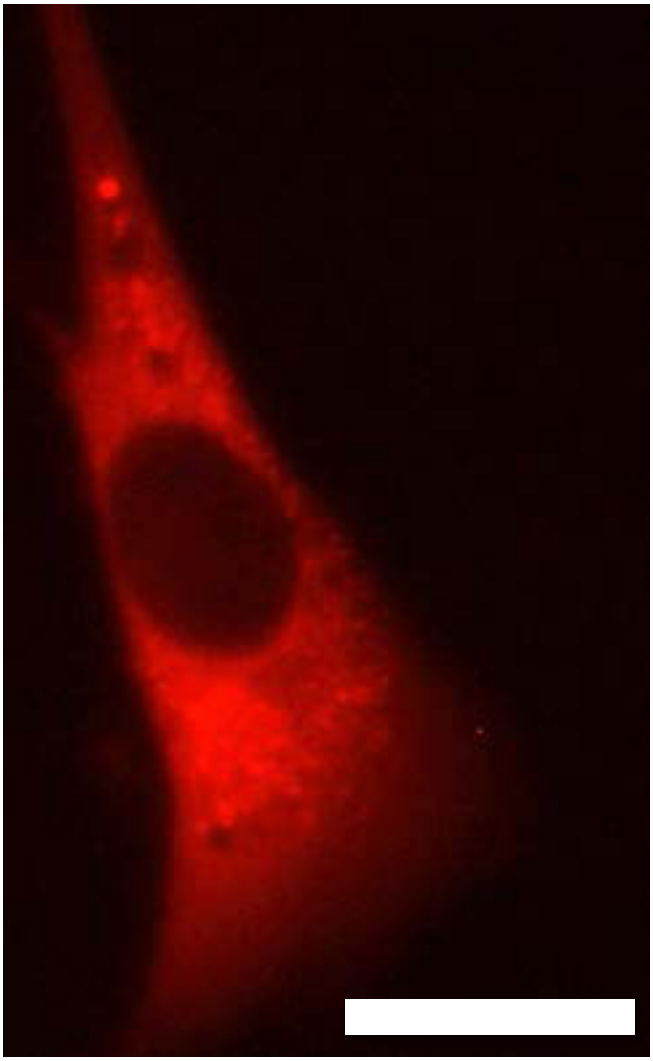


Figure 2.13: A HEK-293FT cell transfected with wild-type KBTBD13/DsRed. The protein appears to localise around the nucleus and is dispersed in the cytoplasm, with intracellular aggregates being seen at higher expression levels. Scale bar represents 50 microns.

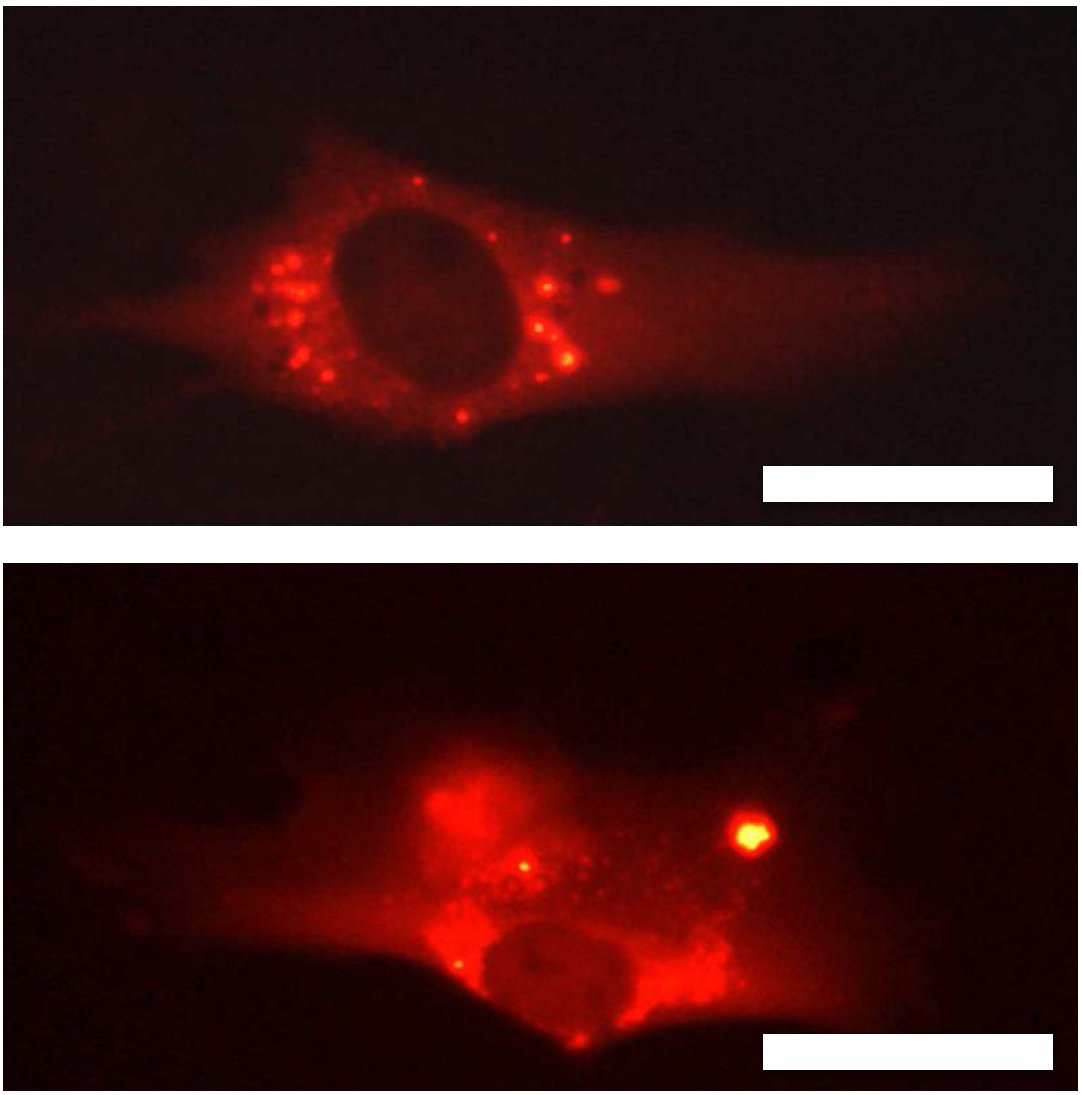


Figure 2.14: C2C12 myoblasts transfected with wild-type KBTBD13/DsRed. The protein appears to localise around the nucleus and is dispersed in the cytoplasm with low expression levels, with aggregations being seen at higher expression levels. Scale bar represents 50 microns.

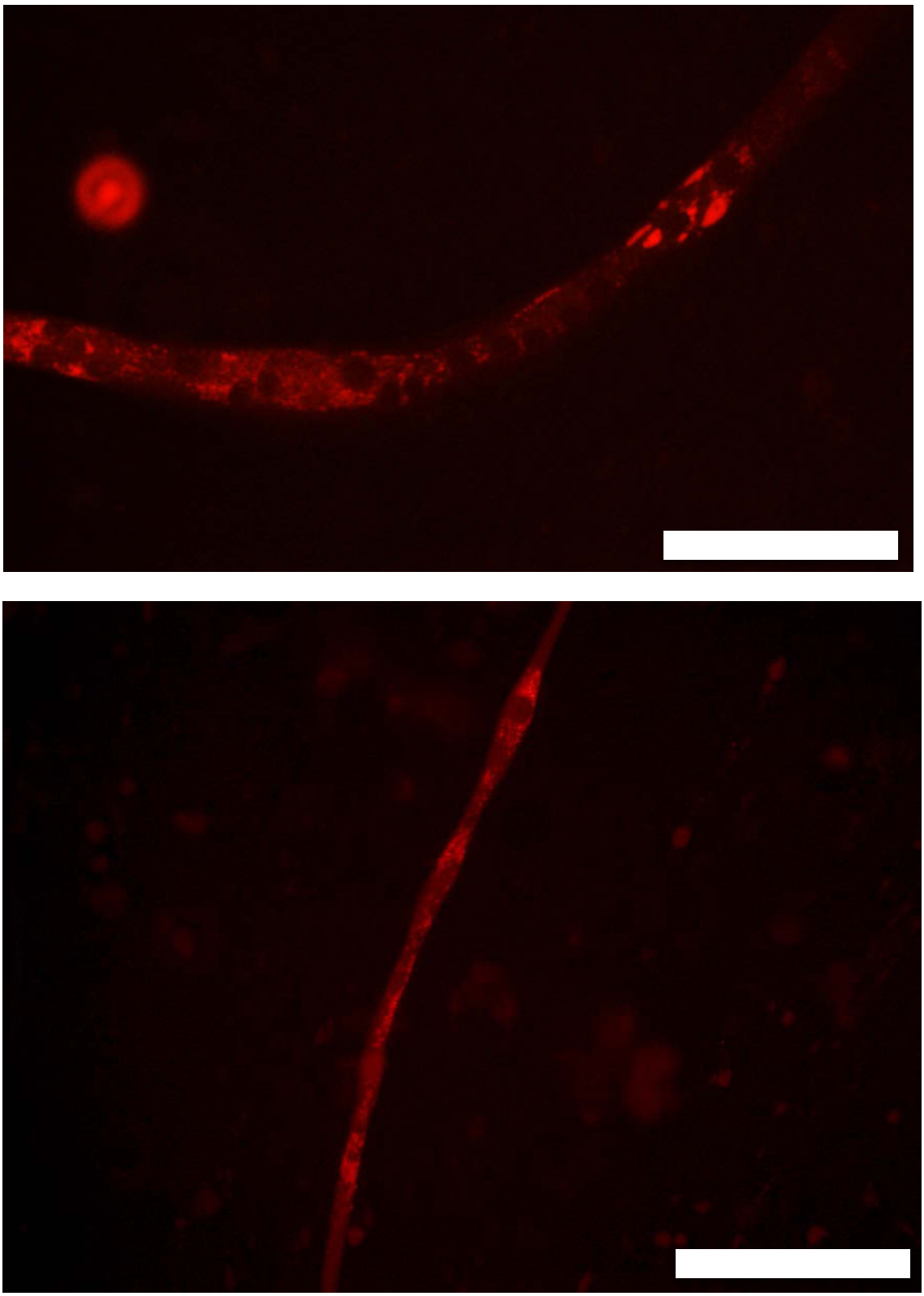


Figure 2.15: Differentiated C2C12 myotubes transfected with wild-type KBTBD13/DsRed. The localisation of the protein does not appear to change between the myoblast and myotube stages of development. Scale bar represents 200 microns.

2.4.8 Polyclonal antibody production

Two rabbit polyclonal antibodies were produced, designated 33D and 79A. Western blot analysis against protein expressed in bacterial and mammalian cell systems found that the produced antibodies were immunoreactive against KBTBD13. However, there were additional bands in the two positive control lanes, and bands of ~ 60 and 70 kDa were seen in the untransfected HEK-293FT lane of the western blot – larger than the expected ~49kDa of endogenous KBTBD13. Furthermore, the antibody was reactive to much higher molecular weight products in the human control muscle lysates. These results can be seen in Figure 2.16.

The non-specificity of these results precluded any further work on human muscle biopsies, which had been planned to be the next step.

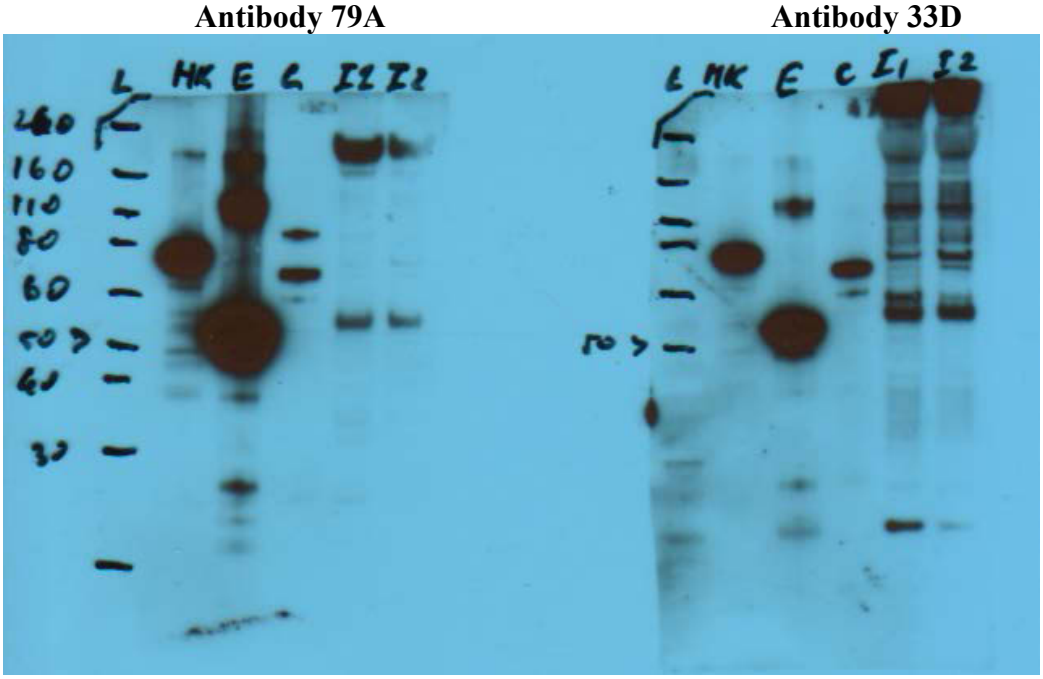


Figure 2.16: Western blots of two antibodies produced against KBTBD13. Molecular weights of markers are labelled on the left side of the L (ladder) lane. Non-specific binding is seen in both antibody 79A and 33D.

Lanes labelled HK are lysates of HEK-293FT cells expressing fluorescently tagged KBTBD13 protein; E lanes are KBTBD13-expressing E. coli lysates; C lanes are untransfected HEK-293FT lysates, and lanes I1 and I2 are human control muscle lyates.

2.5 Discussion

The work described in this Chapter was the first identification of mutations in the gene *KBTBD13*. The mutations were associated with core-rod disease and nemaline myopathy with a unique muscle slowness (NEM6, OMIM #609273). Three separate mutations have been identified in four different families and a single isolated case.

2.5.1 Mutation prevalence

A common mutation in a shared microsatellite region was found in the Australian-Belgian, Australian-Dutch and Dutch families, which are not known to be directly related. This suggested a founder effect in these three families from the neighbouring Belgium and the Netherlands which was confirmed by haplotyping. This particular phenotype has been rarely described in the literature, and recent reviews have only touched upon the genetics of NEM6.[102] No further mutations within KBTBD13 have been published as of the writing of this thesis.

2.5.2 Relationship to known kelch protein mutations and possible pathobiology

Although KBTBD13 protein does not display major sequence homology to KLHL9, both are BTB-Kelch proteins known to be implicated in muscle disease, mutations causing distal myopathy in the case of *KLHL9* mutations. Sambuughin *et al.,* found that KBTBD13 is an adapter protein for the E3-cullin ubiquitin ligase complex by expressing the BTB domain without the kelch domain.[103] As such, the hypothesis could be made that mutations within the Kelch domain, which has protein binding among its functions[104] would cause inefficiencies in protein degradation due to inefficient or negligible target protein binding, and consequent aggregation of proteins that would otherwise be destined for degradation.

The muscle 'slowness' unique to this phenotype of core-rod disease could also imply that a defect in calcium handling is involved as part of the pathobiology. Before the identification of *KBTBD13* as a disease gene causative of core-rod disease, only the ryanodine receptor (*RYR1*) and selenoprotein N1 (*SEPN1*) were known core-rod disease genes. Both of these genes code for proteins that are involved with calcium handling in muscle.[105; 106]

KBTBD13 protein has successfully been expressed in two mammalian cell lines, and was found to be a cytoplasmic protein. Further work by the Goldfarb group confirmed this, and also found that the BTB domain of KBTBD13 interacts with the E3-cullin ubiquitin ligase complex.[95; 103] A polyclonal antibody was designed against a portion of the KBTBD13 protein, but was judged to be insufficiently specific to KBTBD13 protein.

A hypothesis that may be put forward is that KBTBD13 acts as an adapter targeting structural proteins involved with calcium handling for ubiquitination and subsequent degradation. Gradual accumulation of non-functional or damaged proteins within the myofibril could explain the comparatively late age of onset compared to RYR1 and SEPN1 core-rod disease, and the slow progression of weakness.

Since the publication of this work, I have also contributed to the identification of mutations in two further kelch protein genes – *KLHL40*[107] and *KLHL41*[108]. Mutations in both of these genes cause severe recessive foetal akinesia with nemaline rods. Garg *et al.*, 2014 determined that KLHL40 protein localises to the I-band and A-band of the sarcomere, and binds to nebulin, promoting the stability of nebulin and leiomodin 3, and blocks leiomodin 3 ubiquitination.[109] Zhang *et al.,* and Gupta *et al.* found that KLHL41

is a known interaction partner of the Cul3 ubiquitin ligase complex and interacts with sarcomeric thin filament proteins.[108; 110] The identification of a second BTB-Kelch myopathy disease gene implicated in the Cul3 ubiquitin ligase pathway gives weight to the hypothesis that KBTBD13 targets proteins for ubiquitination and subsequent degradation. However, the targeted proteins still remain unknown.

Further investigations could be undertaken by expression of soluble KBTBD13 propeller domain, and using this in a pull-down assay to identify binding partners within skeletal muscle, or using a third-party company to conduct a yeast two-hybrid screen against known muscle proteins. The difficulties in expressing soluble KBTBD13 protein, and creating a specific antibody were reiterated in a personal communication from Coen Ottenheijm to Prof. Nigel Laing; stating that they could not express soluble KBTBD13, or produce an antibody to it. If eventually an antibody is made, or soluble KBTBD13 protein is expressed, experiments could be conducted to assess whether the mutant proteins decrease binding affinity to target substrates.

Chapter 3

Construction and implementation of next generation sequencing bioinformatic pipelines

3.1 Summary

This chapter details the workflow of the next generation sequencing used in this thesis and the bioinformatics processing pipelines that I developed over the course of the thesis. The bioinformatic pipelines described in this chapter have contributed to date to the publication of five papers.

3.2 Introduction

The major foci of my thesis were the application of next generation sequencing (NGS) to: 1) the discovery of novel disease genes, 2) expanding the phenotypes of known disease genes, and 3) translating NGS into molecular diagnostics.

3.2.1 Next generation sequencing and bioinformatics

Due to the large amounts of data produced by next generation sequencers, specialised data analysis tools had to be developed to handle the data. These tools come under the general heading of bioinformatics. Bioinformatics is a broad, interdisciplinary field most simply described as the intersection between biology and information technology.[111]

Multiple bioinformatic programs are needed to process the raw data from a sequencer into a useable format that is easily understandable and useable by genetics researchers and clinicians. These multiple bioinformatics programs are used sequentially in constructed 'bioinformatic pipelines' for the analysis of next generation sequencing data.

Dedicated file formats are needed for efficient storage and handling of the large amounts of data generated by next generation sequencing, in addition to the specialised bioinformatics analysis software.

The various sequencing platforms need different methods of analysis, due to the differences in sequencing chemistry, base calling and machine output. However, there has been significant standardisation of the intermediate output file formats from these NGS sequencing platforms.[112]

The primary NGS platforms used in this Thesis were the Life Technologies platforms at the LotteryWest State Biomedical Facility Genomics: The SOLiD 4 and 5500XL sequencers, and the Ion Proton sequencers. In addition to these, the first individual sequenced was sequenced on an Illumina HiSeq platform, by a commercial supplier of next generation sequencing: the Australian Genome Research Facility.

3.3 Elements of next generation sequencing bioinformatics pipelines

A flow diagram of a bioinformatics pipeline, using the example of exome sequencing, as applied to disease gene discovery, is shown in Figure 3.1. The pipeline is divided into 11 elements and each, as performed in this Thesis will be described.

3.3.1 DNA quality

The use of high quality DNA in NGS is important. If input genomic DNA is of insufficient quality, the quality of the resultant fragment library and resultant sequencing data cannot be guaranteed to be of a standard adequate for analysis. Multiple factors contribute to the measure of DNA quality, but two important measures are total DNA concentration and the level of fragmentation in the DNA sample.

3.3.2 Library preparation

Library preparation is an integral step of NGS. It is the process where input DNA is prepared for sequencing. Because of differences in technology and chemistry, the exact process differs between the sequencing machines used. However, an overview of the common steps involved in library preparation for the sequencers used in this thesis is possible.

1. DNA is fragmented. This can be done by a number of methods, including enzymatic shearing or sonication. The fragment size depends on the type of library being made, and the sequencing technology used. DNA to be exome sequenced on the SOLiD sequencer was fragmented into ~150bp fragments, whereas DNA to be sequenced on the Ion Proton was fragmented into ~200bp fragments.
2. DNA fragments are end-repaired (nucleotide overhangs at the end of fragments are converted into blunt ends), and sequencing adapters are ligated on. The sequencing adapters may be barcoded with specific nucleotide sequences to enable multiplex sequencing of, for example, DNA samples from different patients on the same sequencing run.

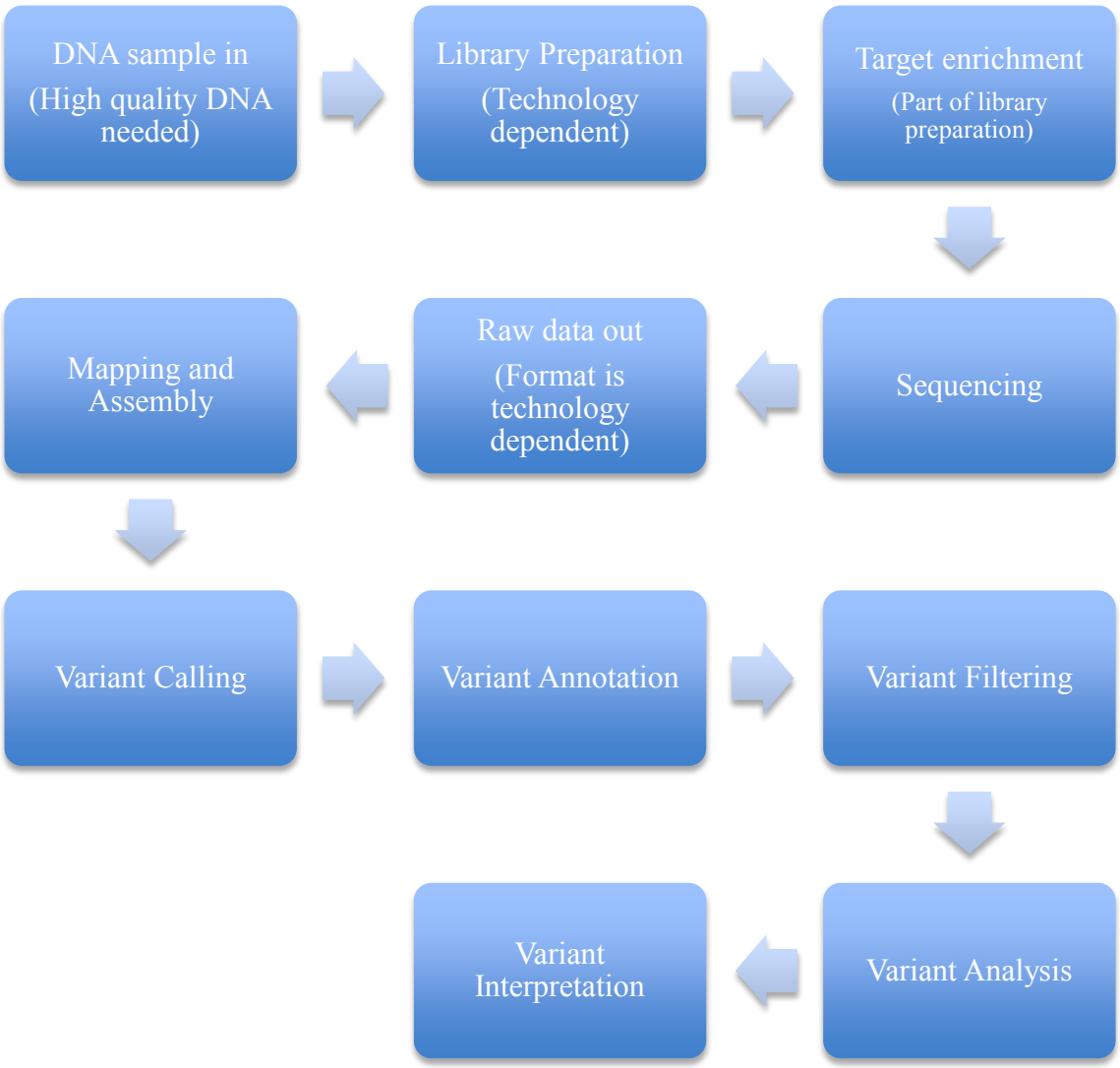


Figure 3.1: Flow diagram of an NGS pipeline from DNA sample input to useable variant data out.

3.3.3 Target enrichment

Target enrichment is performed for many NGS applications, except in whole genome sequencing, or if you have physically isolated a chromosome.[113]

For exomic and sub-exomic sequencing, as opposed to whole genome sequencing, regions of interest within the genome are enriched for in the final sequencing library by a variety of methods. Hybridisation capture was the enrichment method used in this thesis, where a fragment library was denatured into single-stranded DNA before biotin-tagged oligonucleotides were hybridised to the DNA fragments in the prepared library. After hybridisation, streptavidin-coated magnetic beads were used to pull down targeted fragments, unbound DNA was washed away and a final amplification and subsequent quality control of the enriched library was performed.

Another method that was not used in this thesis is enrichment by amplification. A commercial version of this technology is the AmpliSeq technology from Life Technologies. The technology uses a set of massively-multiplex PCR primer sets to enrich and amplify regions of interest in a single step. There is an exome sequencing kit available, although it was not used in the course of this thesis. The AmpliSeq technology has been validated clinically on the Ion Torrent Personal Genome Machine NGS system for screening of cancer hotspots, and identification of mutations in patients with long QT syndrome.[114; 115]

3.3.4 Next generation sequencing

Three different next generation sequencing methods were used in this Thesis: 1) Illumina, 2) SOLiD and 3) Ion Proton.

The chemistry involved in Illumina sequencers has been covered in detail in Chapter 1. A single individual was sequenced using this technology.

The SOLiD series of NGS machines (SOLiD 3, 4, 5500 and 5500XL) work on a unique ligation-based chemistry, with each extension adding a fluorescently tagged eight base probe to an initial sequencing primer, of which three bases are cleaved off before the next ligation. This process continues for a set number of ligations determined by the desired read length (50 or 75 bases in the forwards direction, 50 bases in the reverse). After this, the synthesised strand is denatured off the template strand and another ligation run begins. The next sequencing primer is offset by a single base from the previous one. In conjunction with a two-base encoding method inherent in the ligation probes, 5 rounds of sequencing are sufficient to sequence every base in the read twice (Life Technologies). A graphical overview of this process can be seen in Figure 3.2.

The Ion Proton, also from Life Technologies is a bead-based sequencing by synthesis platform; however, it does not use fluorescently-tagged or chemiluminescent probes for base encoding. Instead, it detects the release of a hydrogen ion upon base incorporation by the use of a proprietary semiconductor-sequencing chip. Sequencing is done in flows; each nucleotide is flowed across the sequencing chip individually and nucleotide incorporation is detected by the voltage change in the well. The magnitude of voltage change scales with the number of nucleotides incorporated.[74] Currently, the average read length produced is 250 bases, with plans to lengthen the reads produced through improved chemistry (personal communication, A/Professor Richard Allcock).

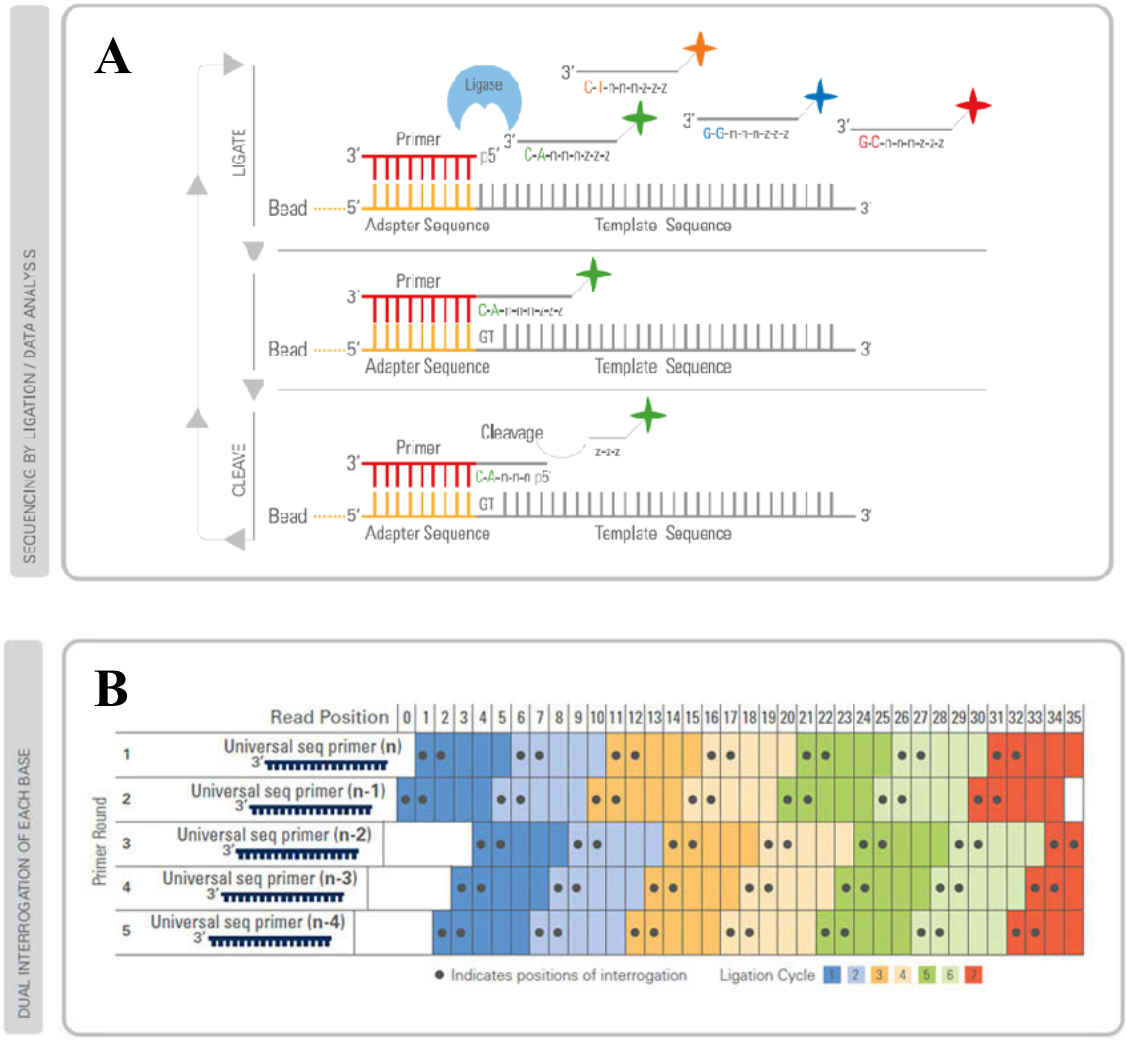


Figure 3.2: A) Process of SOLiD sequencing ligation chemistry, with B) diagram of how 2-base encoding covers each base twice after 5 rounds of sequencing. (Figures were taken from the Life Technologies website (http://www.lifetechnologies.com).

3.3.5 Raw data outputs

The raw data output from some sequencing machines is proprietary, and requires specially tailored bioinformatic tools to process. One example is the raw output from the SOLiD 5500XL machine, which works in what is termed by Life Technologies as 'colourspace'. This is data encoded by the proprietary di-base encoding inherent to the sequencing and contains additional base call quality information.[116] The Illumina sequencers and Ion Proton machines can have their data exported in what is known as '.fastq' format, developed by the Wellcome Trust Sanger Institute. The data files contain the raw sequence reads with attached encoded quality information.[117] However, care must be taken in the assembly, mapping and trimming of these data, as the range of .fastq base quality calls differ between the Ion Proton and Illumina machines[118; 119].

3.3.6 Mapping and assembly

The first part of the analysis of the raw next generation sequencing data is the mapping and assembly of contigs (alignments) from huge numbers of short reads. After mapping and assembly of the reads produced by a sequencer, variants may then be called from the assembled contigs. To this end, a large number of dedicated mapping and assembly algorithms and variant calling algorithms have been developed.

After the raw sequence output is obtained, it is transferred into a separate processing computer for mapping and assembly. In the case of data from the SOLiD sequencers, a compute cluster was used. Data from the Ion Proton machines were processed on a separate processing computer supplied by Life Technologies. The NGS data can be processed using manufacturer-specific, or third party tools. After mapping and assembly, an output file in '.bam' (binary alignment/map) or '.sam' (sequence alignment/map) is produced.[120] The output file generally contains all mapped reads with

attached mapping quality information (which is generated by the mapping and assembly algorithm in each pipeline). It can contain additional data, dependent on the type of NGS machine that generated the raw data.[120] Optimising the mapping parameters used in creating these files was outside of the scope of the work in this thesis.

Analyses of data from ABI SOLiD 4 and 5500XL machines used a mapping algorithm integrated into the ABI LifeScope package, due to the unique sequencing chemistry of the SOLiD sequencers.

The principal sequencing platform used in the latter part of my thesis to generate NGS data was the Life Technologies Proton sequencer. The alignment and variant calling package Torrent Suite is supplied with the Ion Proton sequencer.

3.3.7 Variant calling

Following mapping and assembly, the raw .bam or .sam file is used in the variant calling step, where differences in the sequence compared to a reference genome are detected. Again, there is a choice between using proprietary and third party programs to call variant data. After variant calling, the output data is most often processed into a '.vcf' (variant call format) file. This format was created and standardised by the 1000 genomes project to create a flexible, powerful database file able to contain variant call data from a single individual, all the way up to an entire population.[121]

The variant caller used in the analysis of SOLiD data is DiBayes[122], which is integrated into the LifeScope software package. This caller takes into account the two-base encoding and error-correction data generated by the sequencer when it is calling single nucleotide polymorphisms (SNPs), ostensibly leading to higher quality calls and

increased call sensitivity.[123] The LifeScope package uses a separate insertion and deletion (indel) caller to call small-scale insertions and deletions.

The Torrent Suite program is also used to call variants from the Ion Proton data, in addition to the mapping and assembly duties it performs.

3.3.8 Variant annotation

Variant annotation is the process of adding auxiliary data e.g. gene, transcript and protein information to a putative variant.[124] Variant annotations leverage data from external databases to add valuable additional information to a raw variant call.

3.3.9 Variant filtering

An additional important step in analysis of NGS data is variant filtering. Depending on the type of sequencing performed, variants can number from a few thousand to as many as three million.[124; 125] When looking for disease-causing mutations, filtering out extraneous variants is an essential step. One way that this filtering can be done is by comparison of the variant output with existing variant databases, and excluding variants that occur above a certain percentage in a normal population.

3.3.10 Variant analysis

Variants that pass filtering must be further analysed in order to determine whether they are the cause of disease.

In a representative example of an “average” exome-sequencing run from the SOLiD 5500XL machine, the initial unfiltered number of variants is approximately 30,000. After the rounds of filtering and annotation described previously, the average number of

variants left in the final variant file are approximately 600. As large a change as this is, there are still around 600 variants left as candidates.

3.3.11 Variant interpretation

Interpretation and identification of possible causative variants is more difficult when searching for novel disease genes, as a simple genotype-phenotype correlation is not available. Therefore, a candidate gene approach must be used, where variants in genes are considered for postulated deleterious effect on the protein product, and whether mutations in that protein product are a likely cause of the disease in the patient. Bioinformatic tools that aid this examination of variants in candidate disease genes and can associate them with known disease phenotypes include Exomiser and Endeavour.[126; 127]

3.4.1 Complete Pipelines developed

The first step in the analysis pipelines of the NGS data generated during the course of this thesis used specialised platform-specific mapping and variant calling software.

The customised constructed pipelines were used after the mapping and variant calling stage, at the annotation and filtering steps. The ANNOVAR[128] software was ultimately used as the primary tool for these analyses.

A number of analysis pipelines were created over the course of this thesis, each one superseding the previous one, with data from patients without a confirmed mutation being reanalysed each time the pipeline was upgraded. Comprehensive data on the degree of improvement were not available.

3.4.2 Initial and intermediate pipelines

The first NGS data available to me for development of a bioinformatics pipeline was read data from a single patient, sequenced by the Australian Genome Research Facility using the Illumina platform. Thus, a pipeline was developed that was able to map, assemble, call and annotate variants from raw read Illumina data. Initial assembly of raw read data from fastq format was done using the BWA software suite,[129] mapping against the hg19 build of the human reference genome.

Subsequent variant calling and annotation were performed using the SAMtools[130] and SVA[131] software packages, respectively. This pipeline was also used initially, in a modified form, to analyse data from the SOLiD 4 sequencer at the LotteryWest State Biomedical Facility Genomics. The annotation databases used were the Ensembl transcript database for gene annotation, and the NCBI dbSNP132 SNP database[132; 133] for variant filtering. A flowchart of this pipeline can be seen in Figure 3.3.

When the results from this pipeline proved to be unsatisfactory in variant calling accuracy, annotation and filtering efficacy, computational efficiency and ease of use, it was decided to change from the use of SAMtools as a variant caller and SVA as an annotation utility. The Genome Analysis Tool Kit (GATK)[134] was chosen to replace the use of SAMtools for variant calling,

Shortly after this time, the BioScope software suite, supplied with the SOLiD 4 sequencer, was superseded by the LifeScope suite, which offered increased variant call accuracy for SOLiD data. Therefore, the data from the SOLiD sequencers was not re-mapped or re-called, but simply fed from the LifeScope suite into the annotation software. Annotation of variants from both of these pipelines was then performed using

the ANNOVAR software, using the Ensembl gene annotation database, and the dbSNP132 variant database. Variants present within the dbSNP132 database at a frequency of 5% or greater were excluded. These two pipelines can be seen as a flowchart in Figure 3.4.

As a comparison, a full annotation run of a single exome including format conversion using version 1.2 of SVA took approximately one hour to complete on an eight-core Intel Xeon-based workstation, not including analysis of variants. A full annotation run including variant filtering and initial *in-silico* analysis of the same exome data using ANNOVAR on a similar machine took approximately twenty minutes.

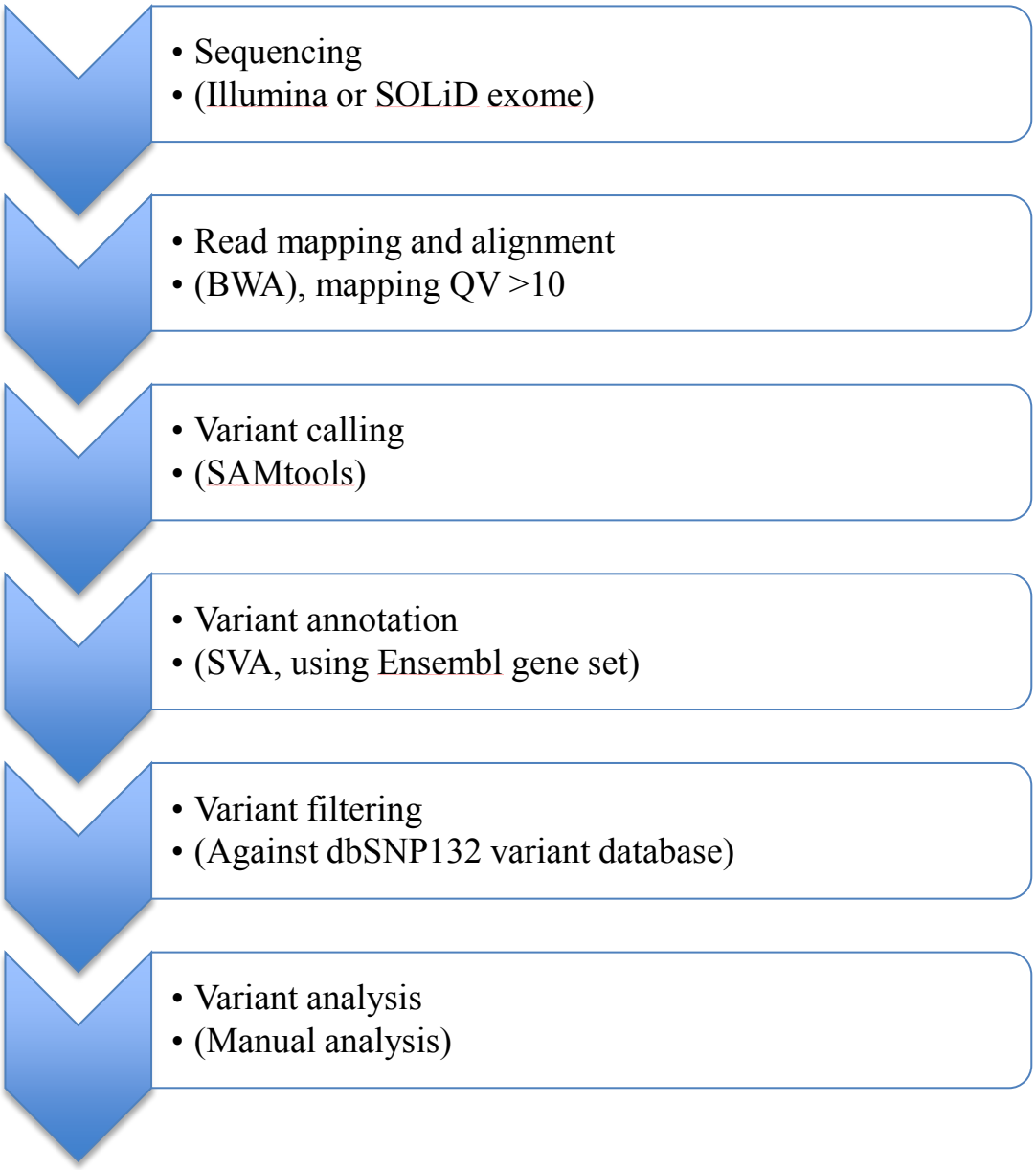


Figure 3.3: Flow diagram of the first bioinformatic pipeline constructed, initially used to analyse data from a single exome from an Illlumina sequencer.

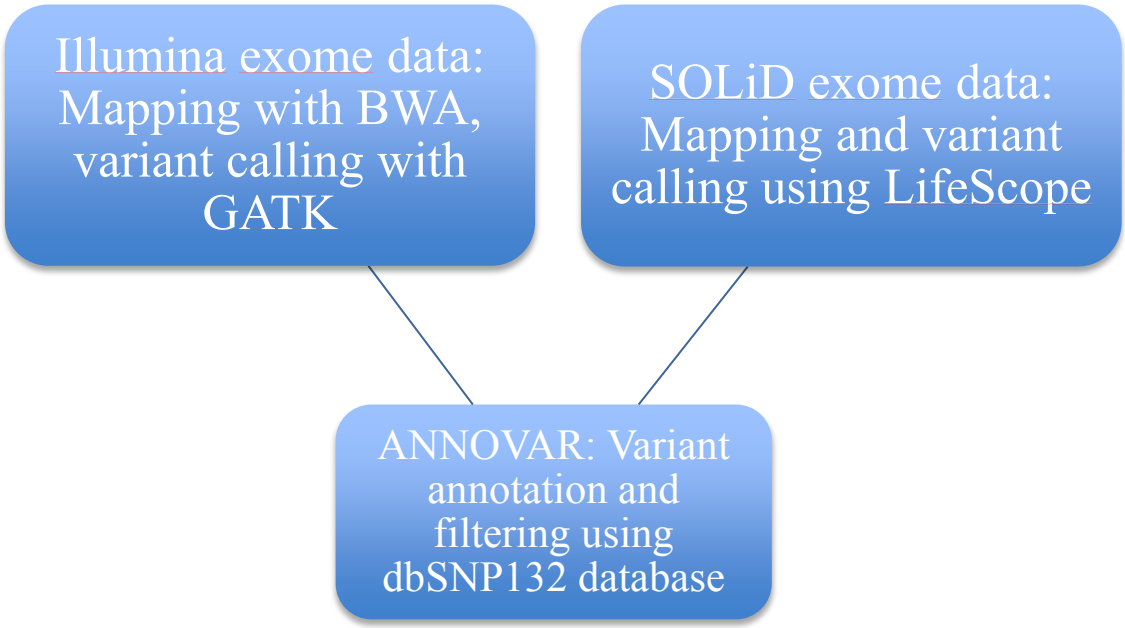


Figure 3.4: Graphical overview of the intermediate pipeline produced to analyse both Illumina and SOLiD exome data. Variants present within the dbSNP132 database at a frequency of 5% or greater were excluded.

3.4.3 ANNOVAR-based pipelines

Using ANNOVAR, four iterations of gene annotation databases were generated. In pipeline V1, the gene annotation database was the RefSeq gene set from UCSC, which was found to have some erroneous transcript annotation. In pipeline V2, the gene annotation database was changed to the Encode GencodeV12 comprehensive transcript annotation set.[135] For pipeline V3, the GencodeV12 database was updated to the GencodeV14 database and then to GencodeV19 in pipeline V4 when the GenecodeV19 data were released.

For the filtering of common single nucleotide variants (SNPs) and indels, two databases were used as filters in the final pipeline V4: 1) the NCBI dbSNP database,[1] dbSNP138Common, and 2) the combined 1000 genomes May 2012 release.[136] These databases were used to filter out variants on the basis of minor allele frequency, where variants with a minor allele frequency of more than 1% were excluded. A third database was incorporated in to the annotation files: the NHLBI GO Exome Sequencing Project (ESP), or Exome Variant Server database.[137] The Exome Variant Server database is known to contain data from patient cohorts, these data were not used to filter out variants in an automated way; instead variants were assessed manually, but still on the basis of their frequency, with a cutoff of 1% used in my analyses.

Modifications were made to the included scripts from the ANNOVAR suite, so that the constructed pipeline could be used in a completely platform-agnostic manner. The final pipeline has been used to process data from Illumina HiSeq, SOLiD 5500XL, Ion Torrent PGM and Ion Proton sequencers. An overview of the V4 pipeline used to analyse both both exome and targeted resequencing data analysis can be seen in Figure 3.5.

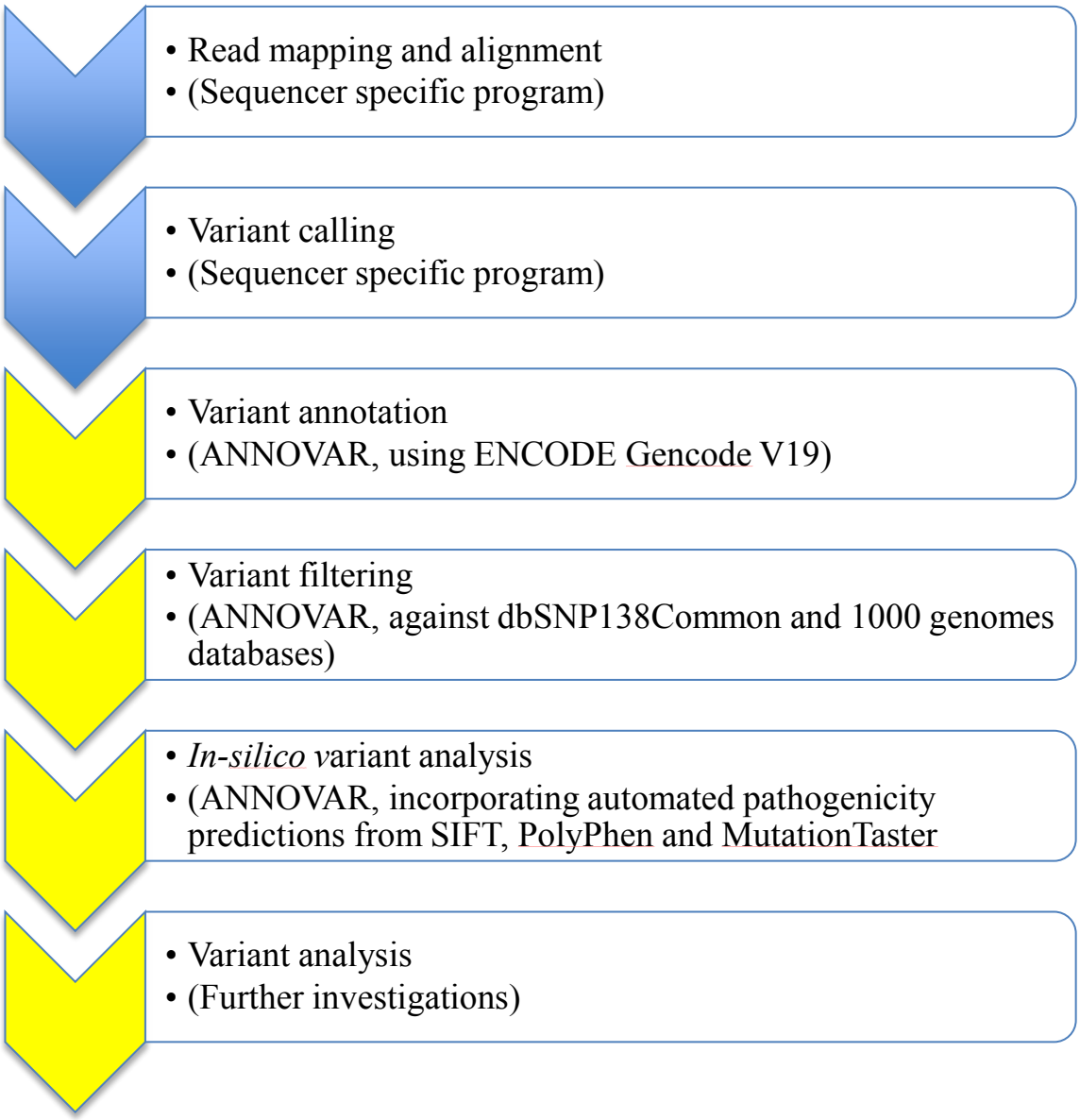


Figure 3.5: Flow diagram of the latest version (V4) of the ANNOVAR variant annotation and filtering pipeline. Chevrons highlighted in yellow are steps performed by the ANNOVAR suite, or the assigned project analyst.

3.4.4 Additional analyses

3.4.4.1 Copy number variant calling

Copy number variant (CNV) calling was implemented using the FishingCNV package.[138] This program was coded to detect copy number variants from exome and targeted resequencing data. The program works by normalising the number of mapped reads per kilobase in a moving window across the genome to create input files, and then comparing these against a normalised set of controls.

For exomes sequenced on the 5500XL, a set of 54 controls were used for the CNV analysis. For the set of targeted resequencing samples sequenced on the Ion Proton, a normalised set of 50 were initially used, then once numbers allowed it, three sets of controls were constructed, a generalised n=200 control set, and two n=100 sex-specific control sets for calling of CNVs on the X-chromosome. Resultant data were then analysed, and visualised in the Integrative Genomics Viewer (IGV)[139].

3.4.4.2 Synonymous variant analysis

An ongoing difficulty with interpreting annotated NGS variant outputs is how to assign importance to synonymous variants. Previous versions of the pipeline used to analyse data in this thesis filtered these synonymous variants out of the final output file, regardless of frequency or exonic position. The final version implemented however, leaves in the annotated output, synonymous variants present in databases at one percent or less.

In order to rank and analyse synonymous variants for possible disease association, a program called SilVA (Silent Variant Analysis using random forests) was used.[140] The program bases its computations predictions on nucleotide conservation, codon usage,

splice sites, splicing enhancers and suppressors, and mRNA folding free energy. An input .vcf file is used, and synonymous variants are filtered and ranked for probable harmfulness according to internal matrices.

3.4.4.3 Common variant filtering

Due to the large amount of data generated by NGS sequencers, a large number of variants, which are common, but not seen in available databases will be present in the output files. These spurious results have many causes, chief of which has been found to be simply variants that while common in the samples sequenced during the course of this thesis, are not present in available online variant databases, or do not have frequency information, and hence are not filtered out by ANNOVAR. The next most common variants are run- and sequencer-specific errors, followed by uncommon alleles that are called as the major allele in the human reference genome.

The above variants were filtered out of the final annotated output by the construction of a laboratory-specific common variant filter file that was added as an extra filtering step to the ANNOVAR analysis pipeline. The addition of the extra filtering step reduces the number of variants in the final output file by on average 400 per exome, and 60 per targeted resequencing run.

3.4.4.4 Combing linkage or exclusion analysis with NGS data to increase the ability to identify causative variants.

If linkage or exclusion data is available for a family, the use of these data as a further variant filtering step can be a powerful addition, restricting the number of candidate variants. If the linkage data is, for example, from a large family, few candidate variants may remain, after combining the linkage data with the NGS data. An initial published

instance of combining positional cloning with exome sequencing to identify a novel disease gene was by Wang *et al.,* (2010). They successfully mapped a novel spinocerebellar ataxia disease gene to chromosome 2 in a Chinese kindred, and subsequently used exome sequencing to identify the novel dominant spinocerebellar ataxia gene *TGM6*.[141] This approach is used in Chapters 4 and 6.

3.4.5 Variant prioritisation

Variants identified by NGS in the course of this thesis were prioritised by whether they fitted the known inheritance pattern of the disease in the family, and if they were in genes known to be associated with the disease phenotype. This was done in the case of families where linkage exclusion did not fully eliminate all known disease genes associated with the phenotype as candidates. When no variants were identified in known disease genes not excluded by linkage analysis, a novel disease gene was hypothesised to be the cause of disease.

3.4.6 Application of the bioinformatics pipeline: Diagnostic exome sequencing example

An example to highlight the power of exome sequencing as applied to diagnostic testing is the identification of the causative mutation from a single patient exome, with suspected inheritance and phenotype data but no further family information.

A single affected patient from a consanguineous family diagnosed with autosomal recessive Charcot-Marie-Tooth disease was exome sequenced using a SOLiD 5500XL sequencer. The initial variant output from LifeScope was a total of 29,667 variants. The variants were then processed, annotated and filtered using an early iteration of the ANNOVAR-based processing pipeline seen in Figure 3.5. After this filtering process, a

total number of 349 variants were left. In-silico variant analysis and interpretation reduced the candidate variants down to 3, from which the homozygous causative mutation in the known autosomal recessive Charcot-Marie-Tooth disease gene *FIG4* (Polyphosphoinositide phosphatase) was identified and confirmed by Sanger sequencing. As a diagnostic exome, this result has not been published.

3.5 Discussion

The V4 pipeline I constructed is comparable to current leading-edge analysis methods. In particular, the software used past variant calling is very similar to the suite of programs used in the recently published FORGE Canada Consortium rare disease gene discovery project.[142] In addition to this, the five disease gene discovery papers published or in press from the laboratory that have used this pipeline, showcase its effectiveness.

However, running a custom bioinformatic analysis pipeline constructed from open-source utilities comes with a set of caveats for the users:

1. The annotation and variant databases used will require constant, regular revision as new versions are published.

2. If third-party annotation software is used, updates and bugfixes to the software will need to be applied regularly.

3. As increasing amounts of data are analysed, larger amounts of storage and computational power will be needed for comparative analysis within the cohort of patients analysed.

4. As a specialised field, it may be difficult to find an appropriately qualified individual to manage, maintain and update a custom analysis pipeline.

5. It may be more efficient in the long run to use commercially available annotation software supplied by the manufacturers of sequencers that are fully integrated into the mapping and variant calling pipeline such as IonReporter (Life Technologies), or purchase and use complete analysis pipelines that are available such as the Bench Lab NGS (Cartagenia) software. (http://www.cartagenia.com/?product=bench-lab-ngs)

Use of the bioinformatics pipelines and publications arising from the use of the pipelines will be described in subsequent chapters.

Chapter 4

Successful application of the bioinformatic pipelines to novel disease gene discovery and expanding the phenotype of known disease genes

4.1 Summary

The bioinformatic pipelines I developed (described in Chapter 3) were implemented as the backbone of next generation sequencing analysis in the Neurogenetic Diseases Laboratory and subsequently have been applied to a number of disease phenotypes. The research on each disease phenotype was led by a senior researcher within the Laboratory.

The greatest success in the application of the bioinformatic pipelines has been in the foetal akinesias, in work led by Assistant Professor Gianina Ravenscroft and encompassing the project of PhD student Ms Emily Todd. The exome sequencing bioinformatic pipeline was used to identify three novel disease genes: Two novel disease genes were identified for severe autosomal recessive foetal akinesia nemaline myopathy: kelch-like family member 40 (*KLHL40*) and kelch-like family member 41 (*KLHL41*). One novel disease gene; striated preferential expressed gene (*SPEG*) was identified for centronuclear myopathy with cardiomyopathy. The analysis of further foetal akinesia cases expanded the genotype phenotype correlations for mutations in glycogen branching enzyme 1 (*GBE1*) and endothelin converting enzyme like 1 (*ECEL1*). In addition to this, a novel cofilin-2 (*CFL2*) mutant causing severe nemaline myopathy was identified, expanding the phenotype of *CFL2* myopathy.

My role was to perform the bioinformatic analysis of all of the data to the point of generating the candidate variant lists. The results have been published and will be reported in greater detail in the PhD thesis of Ms Emily Todd.

4.2 Introduction

The Neurogenetic Diseases Laboratory, in which I studied during my PhD, is a multidisciplinary laboratory focusing on the identification and characterisation of novel neurogenetic disease genes, the development of novel diagnostic tests and translational research in the development of therapies for neuromuscular disorders. Among other high-impact discoveries, the Laboratory was the first to identify mutations in skeletal muscle actin as causative of multiple congenital myopathies including nemaline myopathy.[9]

At the start of my PhD, the Laboratory aimed to apply next-generation sequencing (NGS) to the identification of novel disease genes, as well as expanding the currently known phenotypes of known disease genes. In addition to this, another aim was to apply NGS sequencing to diagnostic testing. (The custom neurogenetic and cardiac disease NGS diagnostic panel developed, is discussed in detail in Chapter 6 of this Thesis).

The bioinformatic pipelines described in Chapter 3 were essential to the identification of 3 novel disease genes by researchers in the Neurogenetic Diseases Laboratory during the course of my thesis, as well as expanding the phenotypes of three known disease genes. Application of the pipeline in projects led by other researchers and students are described in this chapter. Applications of the bioinformatic pipeline in which I was the lead student are in subsequent chapters.

4.3 Materials, methods and results

4.3.1 Identification of *KLHL40* as a disease gene causative of foetal akinesia: (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd).

Two families were investigated in the process of identifying this disease gene. Both families had offspring affected by a very severe nemaline myopathy, resulting in foetal akinesia. The parents in Family one were consanguineous, and had two affected children. Family two had a single affected child from unaffected parents who were not known to be consanguineous.

I performed linkage exclusion analysis on members of the consanguineous family using the MERLIN linkage analysis program on data from a custom 14,514 SNP linkage set based on the Illumina CytoSNP12 beadchip. This analysis excluded all nemaline myopathy and foetal akinesia disease genes known at the time.

DNA from a single affected child from each family was whole-exome sequenced at the Lotterywest State Biomedical Facility Genomics (LSBFG) on a SOLiD 5500XL sequencer (Life Technologies). The resultant exome data from each individual were then mapped to the human genome and had variants called using the LifeScope program (Life Technologies). The resultant mapped data from each child were then processed using the ANNOVAR-based V2 annotation and filtering pipeline described in section 3.4.3 of Chapter 3. Final processed data were then transferred to Ms Emily Todd and A/Prof Gina Ravenscroft for further variant analysis and prioritisation.

Analysis of the processed data identified a homozygous NM_152393:c.1582G>A (p.Glu528Lys) change in Family One, and compound heterozygous changes NM_152393:c.932G>T (p.Arg311Leu) and NM_152393:c.1516A>C (p.Thr506Pro) in

KLHL40. All three variants were confirmed to segregate with disease in their respective families by Ms Emily Todd.

The linkage analysis that I performed on the linkage data from Family One found a positive LOD score of 1.65 at the closest SNP to *KLHL40*. This positive LOD score was not significant linkage, but confirmed that the *KLHL40* locus was not excluded.

Subsequent collaboration with multiple centres around the world ultimately identified 19 *KLHL40* mutants in 28 families, including a founder mutation in *KLHL40* (NM_152393:c.1582G>A) that is responsible in Japan for 28% of cases of foetal akinesia nemaline myopathy.

KLHL40 protein was shown to be absent or greatly reduced in muscle biopsies from the patients. Knock down of *klhl40a and klhl40b* in zebrafish by Dr Rob Bryson-Richardson at Monash University demonstrated gross deformities and disruption of muscle patterning, resulting in severe reduction in movement in the zebrafish.

Publication arising from this work:

Mutations in KLHL40 are a frequent cause of severe autosomal-recessive nemaline myopathy.
Am J Hum Genet. 2013 Jul 11;93(1):6-18.
Ravenscroft G, Miyatake S, Lehtokari VL, Todd EJ, Vornanen P, ***Yau KS****, Hayashi YK, Miyake N, Tsurusaki Y, Doi H, Saitsu H, Osaka H, Yamashita S, Ohya T, Sakamoto Y, Koshimizu E, Imamura S, Yamashita M, Ogata K, Shiina M, Bryson-Richardson RJ, Vaz R, Ceyhan O, Brownstein CA, Swanson LC, Monnot S, Romero NB, Amthor H, Kresoje N, Sivadorai P, Kiraly-Borri C, Haliloglu G, Talim B, Orhan D, Kale G, Charles AK, Fabian VA, Davis MR, Lammens M, Sewry CA, Manzur A, Muntoni F, Clarke NF, North KN, Bertini E, Nevo Y, Willichowski E, Silberg IE, Topaloglu H, Beggs AH, Allcock RJ, Nishino I, Wallgren-Pettersson C, Matsumoto N, Laing NG.*

4.3.2 Identification of *KLHL41* as a disease gene causative of foetal akinesia (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd).

A single family from South Australia was first investigated in the process of identifying this disease gene. The parents in the family were consanguineous, and previously had four unaffected children. A single affected newborn presented with foetal akinesia sequence with nemaline myopathy.

I performed homozygosity mapping on the DNA from the affected newborn using an Illumina CytoSNP12 beadchip. This eliminated all known nemaline myopathy genes except cofilin 2 and nebulin. These two disease genes were subsequently excluded following Sanger sequencing at commercial facilities.

DNA from the affected newborn was subsequently exome sequenced at LSBFG on a SOLiD 5500XL sequencer (Life Technologies). Data were processed as per section 4.3.1, resulting in 453 heterozygous or homozygous variants passing filtering. Application of the homozygosity mapping data to these variants by Ms Emily Todd and A/Prof Gina Ravenscroft further reduced the variant number to seven candidate variants.

Of these seven variants, two variants were skeletal-muscle specific, and the most likely candidate variant was a homozygous eight bp deletion within *KLHL41*: NM_006063: c.1748_1755delAAGGAAAT, causing a change p.Lys583Thrfs*7. This variant was found to segregate with disease in the family by Ms Emily Todd using Sanger sequencing.

Contact with Professor Alan Beggs' group revealed that they had identified mutations in *KLHL41* a few days previously, using exome sequencing on an Illumina HiSeq 2000 machine. A total of seven different *KLHL41* mutations were identified in five families of various ethnic backgrounds.

Antisense morpholino knockdown of the zebrafish isoforms of *KLHL41* (*klhl41a* and *klhl41b)* recapitulated the pathology seen in muscle biopsied from affected individuals, with disorganised myofibrils and nemaline bodies, in addition to Z-line thickening seen under electron microscopy.

Publication arising from this work:

Identification of KLHL41 Mutations Implicates BTB-Kelch-Mediated Ubiquitination as an Alternate Pathway to Myofibrillar Disruption in Nemaline Myopathy.
Am J Hum Genet. 2013 Dec 5;93(6):1108-17.
*Gupta VA, Ravenscroft G, Shaheen R, Todd EJ, Swanson LC, Shiina M, Ogata K, Hsu C, Clarke NF, Darras BT, Farrar MA, Hashem A, Manton ND, Muntoni F, North KN, Sandaradura SA, Nishino I, Hayashi YK, Sewry CA, Thompson EM, **Yau KS**, Brownstein CA, Yu TW, Allcock RJ, Davis MR, Wallgren-Pettersson C, Matsumoto N, Alkuraya FS, Laing NG, Beggs AH.*

4.3.3 Identification of *SPEG* as a disease gene causative of centronuclear myopathy with dilated cardiomyopathy (Lead researchers Dr Gianina Ravenscroft, Prof Nigel Laing)

A Turkish family with no known consanguinity was investigated in the process of identifying this disease gene. The patient was diagnosed with centronuclear myopathy upon biopsy and then with dilated cardiomyopathy at one month of age.

DNA from the Turkish proband was subsequently exome sequenced at LSBFG on a SOLiD 5500XL sequencer (Life Technologies). Data were processed as per section 4.3.1 and the proband was found to be compound heterozygous for a frameshift variant

(NM_005876:c.2915_2916delCCinsA [p.Ala972Aspfs*79]) and a missense mutation (NM_005876:c.8270G>T [p.Gly2757Val]) in *SPEG*.

Collaboration, with Professor Alan Beggs' group taking the lead, resulted in the addition of the Turkish family to their existing cohort of two families with *SPEG* mutations. Professor Beggs' group had previously identified that SPEG protein interacted with myotubularin, the protein product of the known centronuclear myopathy disease gene *MTM1*.

Publication arising from this work:

Am J Hum Genet. 2014 Aug 7;95(2):218-26.
SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy.
Agrawal PB, Pierson CR, Joshi M, Liu X, Ravenscroft G, Moghadaszadeh B, Talabere T, Viola M, Swanson LC, Haliloğlu G, Talim B, ***Yau KS****, Allcock RJ, Laing NG, Perrella MA, Beggs AH.*

4.3.4 Expanding the phenotype of disease caused by mutations in *GBE1* (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd)

A non-consanguineous Caucasian family was investigated to identify the cause of disease in three affected children, of which two were monochorionic twins. The first pregnancy was terminated at 26 weeks, with the autopsy showing a foetus with severe symmetrical contractures. Hematoxylin and eosin staining of the gastrocnemius muscle revealed a dystrophic pattern with increased fibrosis and replacement with fatty tissue, ultrastructural analysis of biceps found disorganised myofibres with fragmented myofibrils. Sanger sequencing excluded the subunits of the acetylcholine receptor (*CHRNA1*, *CHRND* and *CHRNG*) and the commonly mutated exon of *DOK7*.

I performed linkage analysis on DNA from both parents and the two affected foetuses, and bioinformatic analysis of exome sequencing data from two affected foetuses as per section 4.3.1. After application of linkage exclusion results to the combined exome

sequencing data, only 16 candidate genes remained in non-excluded regions. The only unexcluded known foetal akinesia gene was *GBE1*, with a known essential splice site mutation NM_000158:c.691+2T>C (intron 5, rs192044702) and a novel missense mutation in exon 7 NM_000158:c.956A>G (p.His319Arg).

This genetic diagnosis prompted the reinvestigation of the histopathological findings and clinical diagnosis, resulting in a clinical re-diagnosis in line with the genetic diagnosis of a severe case of glycogen storage disease 4.

Publication arising from this work:

Neuromuscul Disord. 2013 Feb;23(2):165-9.
Whole exome sequencing in foetal akinesia expands the genotype-phenotype spectrum of GBE1 glycogen storage disease mutations.
*Ravenscroft G, Thompson EM, Todd EJ, **Yau KS**, Kresoje N, Sivadorai P, Friend K, Riley K, Manton ND, Blumbergs P, Fietz M, Duff RM, Davis MR, Allcock RJ, Laing NG.*

4.3.5 Expanding the phenotype of disease caused by mutations in *ECEL1* (Lead researchers Dr Gianina Ravenscroft, PhD Student Ms Emily Todd)

An Australian family was investigated to identify the molecular cause of disease in their two affected children. The parents were unaffected and not known to be consanguineous. Prominent pterygia and ptosis in the proband lead to consideration of multiple pterygium syndrome in both children, but Sanger sequencing of the *CHRNA* and *CHRNG* genes found no mutations.

Whole exome sequencing was performed as per previous sections and upon subsequent bioinformatic processing as per section 4.3.1, variants were identified in 205 genes. Of these genes, the only known neuromuscular disease genes was *ECEL1*. Compound heterozygous mutations c.1531G>A, (p.Gly511Ser) and a splice-site variant at the intron 12/exon 13 boundary: c.1797-1G>A were identified. Mutations in *ECEL1* are

known to only cause of distal arthrogryposis 5D.[143-145] The proband investigated had the additional features of epicanthal pterygia, a feature previously not associated with distal arthrogryposis 5D.

Publication arising from this work:

Am J Med Genet A. 2014 Jul;164(7):1846-9.
Distal arthrogryposis type 5D with novel clinical features and compound heterozygous mutations in ECEL1.
Barnett CP, Todd EJ, Ong R, Davis MR, Atkinson V, Allcock R, Laing N, Ravenscroft G.

4.3.6 Expanding the phenotype of disease caused by mutations in *CFL2* (Lead researchers Prof Nigel Laing, Mr Royston Ong, *BSc*)

A consanguineous Saudi Arabian family was investigated to identify the molecular cause of disease. Of four children, three were affected, two passing away at 12 and 14 months of age. Pathological diagnosis found nemaline bodies in gomori-trichrome staining.

A single proband from the family was exome sequenced and after bioinformatic filtering as per section 4.3.1, 36 homozygous variants remained. Of these, one was a deletion c.100_103delAAAG, p.Lys34Glnfs*6 in the known nemaline myopathy gene *CFL2*. Subsequent immunostaining found complete absence of CFL2 protein in muscle biopsies from the proband. This is now the third *CFL2* mutation to be described in literature, with a more severe clinical phenotype than in the other two families described where the patients have missense rather than genetic null mutations.[88; 146]

Publication arising from this work:

J Neurol Neurosurg Psychiatry. 2014 Sep;85(9):1058-60.
Novel cofilin-2 (CFL2) four base pair deletion causing nemaline myopathy.
*Ong RW, AlSaman A, Selcen D, Arabshahi A, **Yau KS**, Ravenscroft G, Duff RM, Atkinson V, Allcock RJ, Laing NG.*

4.4 Discussion and conclusions

From these results, the effectiveness in gene discovery of NGS combined with classical methods of linkage analysis and homozygosity mapping, in identifying novel disease genes in small families, can be seen. In many of these families traditional, non-NGS methods of disease gene discovery would most likely have failed or required much greater time and resource commitment.

To successfully identify disease genes, bioinformatic tools must be leveraged effectively to reduce the data output from NGS to a number of candidate variants that is easily processed by an analyst. A well-constructed bioinformatic pipeline with additional supporting tools, including linkage data, can efficiently perform this task, reducing the number of candidate variants before manual examination of variants is required. This is particularly evident in the family where *KLHL41* was identified as a disease gene, as the number of candidate variants was reduced to seven before manual examination.

In addition to the effectiveness in disease gene discovery, NGS diagnostics identify mutations in already known disease genes that would not normally be investigated based on clinical diagnosis. The *GBE1* case, is a good example of such expanding of the phenotypes associated with a known disease genes and so called “diagnosis by sequencing”. The data from exome sequencing of two affected foetuses and linkage exclusion data resulted in re-examination of muscle biopsies by the pathologists involved with the family, and subsequent re-classification of the diagnosis.

Similarly, the *ECEL1* family showed that there can be overlap between the phenotypes of distal arthrogryposis 5D and multiple pterygium syndrome, and *ECEL1* should be

considered as a disease gene when either distal arthrogryposis or multiple pterygium syndrome are diagnosed.

The *CFL2* publication identified only the third family in the world described with a *CFL2* mutation and expanded the severity of the spectrum of nemaline myopathy known to be associated with mutations in *CFL2*.

Chapter 5

Next generation sequencing methods applied to disease gene discovery in hereditary spastic paraplegia and cerebellar ataxia

5.1 Summary

In this chapter the application of NGS to three inherited neuropathy families will be described. One of the families had a clinical diagnosis of spastic paraplegia and the other two families had a clinical diagnosis of cerebellar ataxia.

Candidate disease genes were identified in two of the three families, but further functional characterisation of these genes have been halted until further families with similar phenotypes have been identified.

5.2 Introduction

The inherited neuropathies are a large group of diseases with heterogeneous phenotypes and genetic causes. The inherited (autosomal dominant, recessive, or X-linked) neuropathies can be divided into 3 disparate groups: (1) isolated neuropathy involving peripheral nervous system exclusively; (2) multisystem neuropathy involving both central (brain or spinal cord) and peripheral nervous systems; and (3) neuropathy with multiorgan involvement affecting non-neurologic organs such as skin, kidney, heart, and liver.[147]

This chapter focuses on the application of traditional and next generation gene discovery methods in concert to identify the genetic cause of disease in three inherited neuropathy families.

5.2.1 Hereditary spastic paraplegia

The hereditary spastic paraplegias (HSPs) are a genetically and phenotypically heterogeneous class of disease, characterised by the presence of lower limb weakness and spasticity. The onset of HSP may be subtle, for example manifesting as lower limb stiffness, or uneven wear on the shoes. The phenotypic spectrum of the HSPs can range from early to late-onset and 'pure' to complex, depending on the causative mutation. Pure HSP is defined by progressive spasticity and weakness of the lower limbs, occasionally occurring with sensory disturbances or bladder dysfunction. Complex HSPs can include cognitive impairment, dementia, epilepsy, extrapyramidal disturbances, cerebellar involvement, retinopathy, optic atrophy, deafness, polyneuropathy, or skin lesions in the absence of coexisting disorders .[148-150]

The genetic spectrum of HSPs encompasses autosomal dominant and recessive inheritance as well as X-linked and mitochondrial inheritance.[150; 151] Epidemiological studies of HSP are sparse, but prevalence is estimated at 1.8 cases per 100 000 globally.[152] Currently, 80 genetic loci causative of HSP have been identified, with a total of 62 HSP disease genes identified.[149] Of these, 19 follow an autosomal dominant (AD) pattern of inheritance, 55 an autosomal recessive pattern of inheritance (AR), 5 are X-linked, and one is maternally inherited.[149] These data can be seen in Table 5.1. Of the known HSP disease genes, mutations in spastin are the most frequent cause of familial HSP, causing 37.6% of cases in an Italian cohort.[149] In this cohort, the rest of the known genes accounted for 42.3% of cases, while 20.1% of familial cases remained without a genetic diagnosis. Of the sporadic HSP cases from the same cohort, spastin again made up the majority of diagnoses, causing 18.8% of cases but 51.8% of the sporadic HSP cases in the cohort remained without a genetic diagnosis.[149]

5.2.2 Known pathobiology of HSPs

The identified HSP genes code for proteins in diverse protein families. There is similarity between the protein structure of some HSP proteins, in particular spastin (*SPAST*, SPG4) and paraplegin (*PGN*, SPG7) which share an ATPases Associated with diverse cellular Activities (AAA) domain; and spartin, which shares a conserved 80-amino acid putative microtubule interacting domain with spastin.[148; 153] Some of the proteins mutated in HSP are also binding partners, for example, spastin and atlastin.[154]

The protein products of the HSP genes, despite their diverse functions can be grouped into ten functional networks in cellular function:[155]

1. Autophagy
2. Axonal transport
3. DNA repair and nucleotide metabolism
4. Endosomal functions
5. Endoplasmic reticulum membrane remodelling
6. Endoplasmic-reticulum-associated protein degradation pathway and protein folding
7. Lipid metabolism
8. Mitochondrial function
9. Myelination
10. Neuron development/synapse related

Because of the large number of HSP disease genes identified, a clearer clinical picture of the disease can be constructed using the known genotype-phenotype correlations, leading to a greater understanding of neuronal function, as well as more comprehensive diagnostic protocols, such as the one proposed by Lo Giudice (2014).[149]

Table 5.1: Table of known HSP loci/genes with chromosomal location and mode of inheritance.

Type	Inheritance	Location	Gene
SPG1	X-linked	Xq28	*L1CAM*
SPG2	X-linked	Xq22.2	*PLP1*
SPG3A	AD	14q22.1	*ATL1*
SPG4	AD	2p22.3	*SPAST*
SPG5A	AR	8q12.3	*CYP7B1*
SPG6	AD	15q11.2	*NIPA1*
SPG7	AR	16q25.3	*PGN*
SPG8	AD	8q24.13	*KIAA0196*
SPG9	AD	10q23.3-25.2	Unknown
SPG10	AD	12q13.3	*KIF5A*
SPG11	AR	15q21.1	*KIAA1840*
SPG12	AD	19q13.32	*RTN2*
SPG13	AD	2q33.1	*HSPD1*
SPG14	AR	3q27-q28	Unknown
SPG15	AR	14q24.1	*ZFYVE26*
SPG16	X-linked	Xq11.2	Unknown
SPG17	AD	11q12.3	*BSCL2*
SPG18	AR	8p11.23	*ERLIN2*
SPG19	AD	9q	Unknown
SPG20	AR	13q13.3	*SPG20*
SPG21	AR	15q22.31	*ACP33*
SPG22	X-linked	Xq13.2	*SLC16A2*
SPG23	AR	1q24-q32	Unknown
SPG24	AR	13q14	Unknown
SPG25	AR	6q23-q24.1	Unknown
SPG26	AR	12p11.1-q14	*B4GALNT1*
SPG27	AR	10q22.1-q24.1	Unknown
SPG28	AR	14q22.1	*DDHD1*
SPG29	AD	1p31.1-p21.1	Unknown
SPG30	AR	2q37.3	*KIF1A*
SPG31	AD	2p11.2	*REEP1*
SPG32	AR	14q12-q21	Unknown
SPG33	AD	10q25.2	*ZFYVE27*
SPG34	X-linked	Xq24-q25	Unknown
SPG35	AR	16q23.1	*FA2H*
SPG36	AD	12q23-q24	Unknown
SPG37	AD	8p21.1-q13.3	Unknown
SPG38	AD	4p16-p15	Unknown
SPG39	AR	19p13.2	*PNPLA6*
SPG40	AD	Reserved	Unknown
SPG41	AD	11p14.1-p11.2	Unknown

SPG42	AD	3q25.31	*SLC33A1*
SPG43	AR	19p13.11-q12	*C19orf12*
SPG44	AR	1q42.13	*GJC2*
SPG45	AR	10q25.3-q25.1	Unknown
SPG46	AR	9p13.3	*GBA2*
SPG47	AR	1p13.2	*AP4B1*
SPG48	AR	7p22.1	*KIAA0415*
SPG49	AR	14q32.31	*TECPR2*
SPG50	AR	7q22.1	*AP4M1*
SPG51	AR	15q21.2	*AP4E1*
SPG52	AR	14q12	*AP4S1*
SPG53	AR	8p22	*VPS37A*
SPG54	AR	8p11.23	*DDHD2*
SPG55	AR	12q25.31	*C12orf65*
SPG56	AR	4q25	*CYP2U1*
SPG57	AR	3q12.2	*TFG*
SPG58	AR	17p13.2	*KIF1C*
SPG59	AR	15q21.2	*USP8*
SPG60	AR	3p22.2	*WDR48*
SPG61	AR	16p12.3	*ARL6IP1*
SPG62	AR	10q25.31	*ERLIN1*
SPG63	AR	1p13.3	*AMPD2*
SPG64	AR	10q24.1	*ENTPD1*
SPG65	AR	10q25.32-q25.33	*NT5C2*
SPG66	AR	5q32	*ARSI*
SPG67	AR	2q33.1	*PGAP1*
SPG68	AR	11q13.1	*FLRT1*
SPG69	AR	1q41	*RAB3GAP2*
SPG70	AR	12q13	*MARS*
SPG71	AR	5p13.3	*ZFR*
GAD1 gene	AR	2q31.1	*GAD1*
SPOAN	AR	11q13	Unknown
	AR	5p15.2	*CCT5*
Leigh syndrome	Maternal	Mit	*MT-ATP6*
	AR	19q13.32	*OPA3*
SMALED2	AR	9q22.31	*BICD2*
	AR	19q13.1	*MAG*
SPG72	AR	5q31	*REEP2*
CHS	AR	1q42.3	*LYST*

5.2.3 Inherited spinocerebellar ataxias

The spinocerebellar ataxias (SCAs) are a group of progressive neurodegenerative diseases, characterised by cerebellar ataxia, resulting in unsteady gait, clumsiness and disarthria. This pure cerebellar phenotype is often accompanied by other neurological signs specific to the type of SCA such as ophthalmoplegia, pyramidal and extrapyramidal signs, peripheral neuropathy, and dementia, among others.[156; 157]

5.2.3.1 Dominantly inherited spinocerebellar ataxias

Dominantly inherited spinocerebellar ataxias (SCAs) are a major subset of the cerebellar ataxias. Originally, autosomal dominant SCAs were thought to be caused solely by the expansion of coding CAG polyglutamine repeats in the disease genes, since such repeats were identified in the first 6 SCAs for which the disease gene was identified.[156] However, now expansions in non-coding areas of genes and point mutations are known to cause dominant SCAs. Out of 31 SCA loci, 8 are known to be caused by coding repeat expansions, 4 by non-coding repeat expansions, 10 by non-repeat expansion mutations and 9 have a locus, but no disease gene identified as yet.[156] These data can be seen in Table 5.2.

5.2.3.2 Polyglutamine expansion SCAs

The polyglutamine expansion SCAs are the most well studied group of dominant SCAs, and causing approximately 45% of the dominant SCAs in a European population.[156; 158] The polyglutamine expansion SCAs present with diffuse neurological dysfunction, with cause of death being brainstem failure. On average, the age of onset is in the third or fourth decade of life; however, the age of onset can be influenced by the number of polyglutamine repeats in the mutated gene, with increased repeat number inversely correlated with age of onset. Some types of expansion SCA are more influenced by repeat number than others.[158]

5.2.3.3 Autosomal recessive spinocerebellar ataxias

The autosomal recessive spinocerebellar ataxias (SCAs) encompass a broad spectrum of clinical phenotypes, with over 20 separate sub-types, and 15 known disease genes. (Table 5.3) Of these, Friedrich ataxia is the most common, with a prevalence of 1-2 per 50,000. Friedrich ataxia is also the only recessive ataxia currently known to be caused by repeat expansions.[159]

The autosomal recessive SCAs tend to manifest before 25 years of age; clinical signs are typically balance abnormalities, incoordination, kinetic and postural tremor, and dysarthria with cerebellum, brainstem, and spinocerebellar tract involvement. [160-162]

Table 5.2: The following table lists all currently known autosomal dominant SCA loci and disease genes, where known.

Type	Inheritance	Location	Gene
SCA1	AD	6p23	*ATXN1*
SCA2	AD	12q24	*ATXN2*
SCA3 (MJD)	AD	14q24.3-q31	*ATXN3*
SCA4	AD	16q22.1	Unknown
SCA5	AD	11q13	*SPTBN2*
SCA6	AD	19p13.2	*CACNA1A*
SCA7	AD	3p21.1-p12	*ATXN7*
SCA8	AD	13q21.33	*ATXN8OS, ATXN8*
SCA9	AD	Unknown	Unknown
SCA10	AD	22q13.31	*ATXN10*
SCA11	AD	15q15.2	*TTBK2*
SCA12	AD	5q32	*PPP2R2B*
SCA13	AD	19q13.3-13.4	*KCNC3*
SCA14	AD	19q13.4	*PRKCG*
SCA15/SCA16	AD	3p26.1	*ITPR1*
SCA17/HDL4	AD	6q27	*TBP*
SCA18/SMNA	AD	7q22-q32	Unknown
SCA19/SCA22	AD	1p13.3	*KCND3*
SCA20	AD	11p11.2-q13.3	Unknown
SCA21	AD	7p21.3-p15.1	Unknown
SCA23	AD	20p13	*PDYN*
SCA25	AD	2p21-p15	Unknown
SCA26	AD	19p13.3	Unknown
SCA27	AD	13q34	*FGF14*
SCA28	AD	18p11.22-q11.2	*AFG3L2*
SCA29	AD	3p26	Unknown
SCA30	AD	4q34.3-q35.1	Unknown
SCA31	AD	16q22	*BEAN1, TK2*
SCA35	AD	20p13	*TGM6*
SCA36	AD	20p13	*NOP56*
DRPLA	AD	12p13.31	*ATN1*

Table 5.3: Table of recessive ataxias with cerebellar involvement, data taken from the Neuromuscular Disorders Gene Table 2014 freeze (www.musclegenetable.fr)

Type	Inheritance	Location	Gene
Friedreich ataxia 1	AR	9q13-q21.1	*FXN*
Friedreich ataxia 2	AR	9p23-p11	Unknown
Friedreich ataxia with selective vitamin E deficiency	AR	8q13.1-q13.3	*TTPA*
Infantile-onset spinocerebellar ataxia	AR	10q23.3-q24.3	*PEO1*
Ataxia oculomotor apraxia 1	AR	9p13.3	*APTX*
Ataxia oculomotor Apraxia 2	AR	9q34.13	*SETX*
Autosomal recessive spinocerebellar ataxia, 3	AR	6p23-p21	Unknown
Autosomal recessive spinocerebellar ataxia, 4	AR	1p36	Unknown
Autosomal recessive spinocerebellar ataxia, 5	AR	15q24-q26	Unknown
Autosomal recessive spinocerebellar ataxia, 6	AR	20q11-q13	Unknown
Autosomal recessive spinocerebellar ataxia, 7	AR	11p15	Unknown
Autosomal recessive spinocerebellar ataxia, 8	AR	6q25	*SYNE1*
Autosomal recessive spinocerebellar ataxia, 9 (with ubiquinone deficiency)	AR	1q42.13	*ADCK3*
Spinocerebellar ataxia with axonal neuropathy	AR	14q31-q32	*TDP1*
Marinesco-Sjogren syndrome (cerebellar ataxia with cataract and myopathy)	AR	5q31	*SIL1*
Sensory ataxic neuropathy, disarthria and ophthalmoparesis	AR	15q25	*POLG*
Ataxia telangiectasia	AR	11q22.3	*ATM*
Ataxia telangiectasia-like disorder	AR	11q21	*MRE11A*
Autosomal recessive spastic ataxia of Charlevoix-Saguenay	AR	13q12	*SACS*
Refsum disease-1 (adult)	AR	10q13	*PHYH*
Refsum disease-2 (adult)	AR	6q21-q22	*PEX7*

5.2.4 Aims

This chapter aims to identify the genetic cause of neurodegenerative disorders in three families by combining classical gene discovery methods with next generation sequencing.

5.3 Materials and methods: Families investigated

5.3.1 Family 1: HSP

When first investigated, the studied family consisted of two living generations (Figure 5.1), with two individuals in the first generation confirmed to be affected upon initial examination. The disease is late-onset, with both affected males displaying progressive spasticity from their early 50s. Upon initial examination, individuals III:3, and III:5 through to III:7 were confirmed to be unaffected, whereas individual III:9 was not old enough to be manifesting clinical symptoms and individual III:4 has not been examined by a neurologist.

5.3.2 Family 2: Autosomal dominant SCA

This is a four-generation family with eight affected members (Figure 4.2) with suggestion of anticipation with a younger age of onset with succeeding generations. There is progression of all the symptoms, with the second generation requiring a wheelchair by 70 years, and the third generation by 50 years. The initial symptom in the first generation was gait ataxia in the 40s-50s, followed within 5 years by slurring dysarthria, saccadic interruption of eye movements, intention tremor, and clumsy hand movements. There is evidence of long tract involvement, with increased deep tendon reflexes in the legs and urinary urgency. Progressive atrophy of the cerebellum, brain stem and cerebellar peduncles develop. Molecular testing for the CAG repeat

expansions of SCA 1, 2, 3, 6, 7, 8 and 17 was negative, and all other investigations were within normal ranges.

5.3.3 Family 3: Recessive SCA

This family has two affected brothers out of 8 sibs who presented at age of 55 and 54 years respectively with weakness about the ankles, poor balance and chronic cough. At the time of presentation they had wasting of the calf muscles and with time this progressed to the anterior muscles below the knee and the intrinsic muscles of the hand and forearm. There was weakness of those muscles and more recently fasciculations. Deep tendon reflexes were globally reduced and have disappeared with time. There were cerebellar signs with saccadic interruption of eye movements and a wide-based gait, and, later, dysarthria. The creatine kinase was elevated, between 580 and 1020 (N <195). Nerve conduction studies showed a severe sensorimotor peripheral neuropathy. Muscle biopsy showed numerous atrophic fibres and group atrophy, in keeping with a neurogenic cause. Electron microscopy additionally showed subsarcolemmal accumulation of pleomorphic mitochondria with large electron dense inclusions. MRI of the head showed progressive atrophy of the cerebellum, principally of the vermis, and thinning of the brainstem. No cause for the cough has been found. The probands have six unaffected siblings, unaffected offspring, and their parents are second cousins (Fig. 5.3).

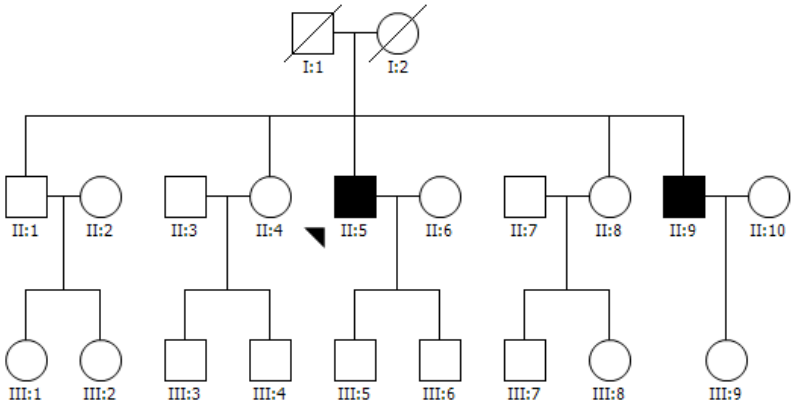


Figure 5.1: Pedigree of family 1, consisting of three generations, with two affected individuals in the second generation. Proband is indicated with an arrow.

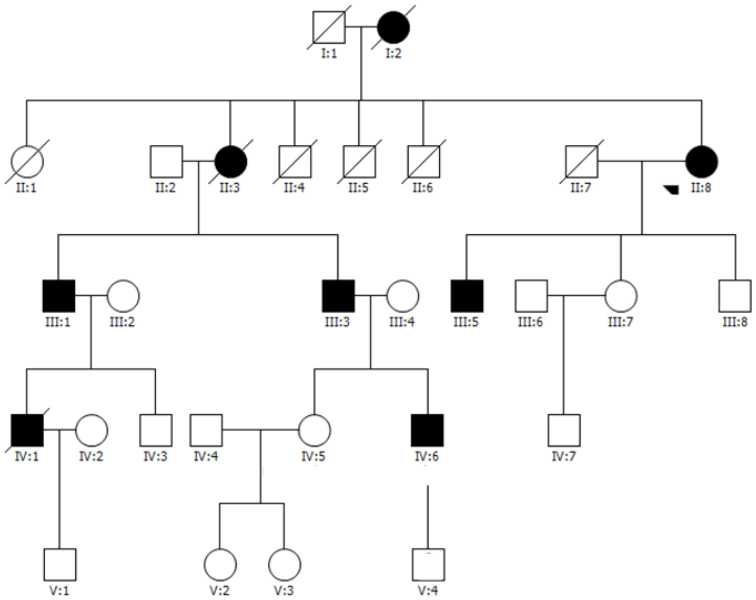


Figure 5.2: Pedigree of family 2, consisting of four extant generations, with six affected individuals. Proband is indicated with an arrow.

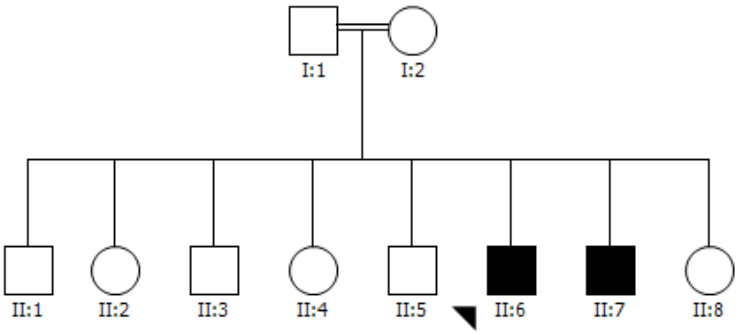


Figure 5.3: Pedigree of family 3, consisting of two generations, with two affected sibs in the second generation. Proband is indicated with an arrow.

5.4 Materials and Methods

5.4.1 Linkage analysis for families 1 and 3

Linkage exclusion on two affected sibs was performed using a custom 14 514 SNP linkage map across the 22 autosomes derived for use from the CytoSNP12 (Illumina) cytogenetics SNP chip. SNP data were analysed across the autosomes using the Merlin[163] linkage suite according to a recessive disease profile, following removal of unlikely genotypes and Mendelian inheritance errors. Regions of the genome with a LOD score of less than -2 were considered to be excluded, as per standard linkage analysis parameters. The regions not excluded by linkage analysis were used as candidate regions in the subsequent next-generation sequencing analysis.

5.4.2 Linkage analysis for family 2

Affecteds-only linkage exclusion analysis was performed using data obtained by Dr Rachael Duff on the Linkage 10k SNP chip from Affymetrix. Initial linkage analyses by Dr Duff did not find significant linkage, but the data were re-examined as more information was obtained on the family.

The data were exported by the Alohomora program suite[164] in the Merlin linkage format, and these data were analysed with the Merlin linkage package[163] using an autosomal dominant disease profile. Regions of the genome with a LOD score of less than -2 were considered to be excluded, as per standard linkage analysis parameters. The regions not excluded by linkage analysis were used as candidate regions in the subsequent next-generation sequencing analysis.

5.4.3 Exome capture and next generation sequencing

Whole exome sequencing was performed on a male affected member (II:5) of family 1 by the Australian Genome Research Facility (AGRF), Brisbane node. Next-generation sequencing was carried out using the Illumina platform.

Whole exome sequencing of an additional affected individual II:9 from family 1, affected individual III:3 from family 2 and affected individuals II:6 and II:7 from family 3 was performed at the Lotterywest State Biomedical Facility Genomics.

Briefly 3 μg of patient DNA was fragmented by sonication using a Covaris and ligated to SOLiD system sequencing adapters. The resulting library was enriched for exomic sequence using the SeqCap EZ Human Exome Library v2.0 exome capture system (Nimblegen, Roche diagnostics) according to manufacturer's specifications. The resultant enriched libraries were sequenced using a SOLiD 5500XL Genetic Analyser (Life Technologies).

5.4.4 Bioinformatic analysis of exome sequencing data

Data from the single affected individual sequenced on the Illumina platform were analysed first using SAMTools and SVA pipeline, and then re-analysed using the GATK and ANNOVAR-based V4 pipeline as described in chapter 3. All other data were analysed using the final ANNOVAR-based pipeline described in chapter 3.

5.4.5 Selection of candidate variants in family 1

Candidate variants from the two affected individuals sequenced were initially prioritised on the basis that the disease was autosomal recessive, given the initial clinical presentation. The two exomes were compared, and shared variants between the two

affected individuals were selected as candidates. After this initial filtering, a linkage exclusion filter was applied to the resultant variant list using the data from section 5.3.4.

5.4.6 Screening of the SCA10 gene by Genescan in family 2

Primer designs and thermocycling protocol used to screen pentanucleotide repeat expansion in the *ATXN10* gene (which was not excluded by linkage analysis) via Genescan (Applied Biosystems) were taken from Cagnoli *et al.,* 2004.

5.4.7 Selection of candidate variants in family 3

Candidate variants were prioritised on the basis that the disease was autosomal recessive, and homozygous and compound heterozygous variants shared between the two affected individuals within unexcluded regions of the genome were prioritised for analysis. As the parents were second cousins, homozygous variants were examined first.

5.4.8 Sanger validation of variants in family 3

Primers for exons containing variants in the genes *TMEM33* and *ZFHX4* were designed to encompass the entire exon containing the variant(s), plus sufficient flanking intronic sequence to produce useable data. The sequences for these primers can be found in Appendix A Table 2.7. All amplicons were amplified using Qiagen HotStar™ Taq polymerase using the TD65 thermocycling protocol and standard Qiagen mix described in Appendix A.

Amplicons were sequenced using the BigDye Terminator mix (ABI) and ABI thermocycling protocol found in Appendix A. Sequencing primers were identical to the PCR amplification primers.

5.5 Family 1 results

5.5.1 Linkage exclusion analysis by SNPs in family 1

Linkage exclusion analysis using an autosomal recessive model, excluded all known recessive HSP disease genes. The majority of recessive HSP disease loci without a known disease gene were also excluded. Incomplete exclusion of the autosomal recessive SPG26 and SPG28 loci were seen, with only a part of the loci excluded. The SPG27, SPG32 and SPG45 loci were not excluded.

As well as these recessive loci, the dominant HSPs SPG3A and SPG33, caused by mutations in atlastin (*ATL1*) and *ZFYVE27*, respectively were not excluded. The dominant loci causing SPG9 and SPG36 were not excluded, and the dominant locus causing SPG19 was incompletely excluded.

5.5.2 Exclusion of known HSP disease genes by exome sequencing

Of the two remaining known HSP disease genes *ATL1* and *ZFYVE27*, *ATL1* was excluded on the basis of the exome sequencing results, as no variants were found within it. Coverage of these genes was 100% in the case of *ATL1*, and all exons save exon 9 of *ZFYVE27* were sequenced fully. Mutations in the *ZFYVE27* gene are currently known only to cause SPG33, which is a pure, autosomal dominantly inherited spastic paraplegia and did not match the phenotype in the patients. Thus, *ZFYVE27* was not considered as a disease gene.

5.5.3 Analysis of variants identified by exome sequencing in family 1

In the initial analysis of the exome sequencing data, after elimination of variants within exclusion regions, all ‘candidate’ variants remaining were either heterozygous, were

artefacts due to reference or annotation errors, or were known variants at a frequency too high in the general population to be considered as disease-causing.

Clinical re-examination of female members of the family found that individuals II:4, II:8 and III:9 were mildly affected. Upon this finding, the linkage exclusion data were re-examined on an X-linked dominant basis, and variants were re-examined on the basis of an X-linked disease. Unluckily, none of the X-chromosome was excluded based on the linkage data.

After re-analysis of the exome sequencing results based upon an X-linked dominant inheritance model, a single NM_005278:c.651C>G (p.N217K) variation in the gene encoding glycoprotein M6B (*GPM6B*) was identified. This variant was not found in the 1000 Genomes, EVS or ExAC databases. GPM6B protein is known to interact with the protein product of proteolipid 1 (*PLP1*), a known HSP disease gene.[165] In the *jimpy*[166] mutant mouse model, GPM6B protein co-localises with mutant PLP1 in the endoplasmic reticulum instead of being transported to the cell membrane.[167]

5.5.4 Sanger validation of variants

Sanger sequencing validated the presence of the NM_005278:c.651C>G (p.N217K) variant in the proband and his affected brother in the hemizygous state, and in two affected sisters in the heterozygous state. Subsequent sequencing of other family members found that a single female (III:9) was heterozygous for the variant, but too young to be clinically manifesting, and all other individuals except individual III:4 sequenced were found to be wild-type. The results of this sequencing can be found in figure 5.5.

Individual III:4, although of age where he should have started to manifest clinical symptoms, would not present for clinical examination, and thus segregation of the variant with disease cannot be confirmed.

Options for further work on this family are severely limited unless individual III:4 presents for clinical examination. The other possibility for progress would be the identification of another family (or families) with a similar phenotype with a variant in the same gene.

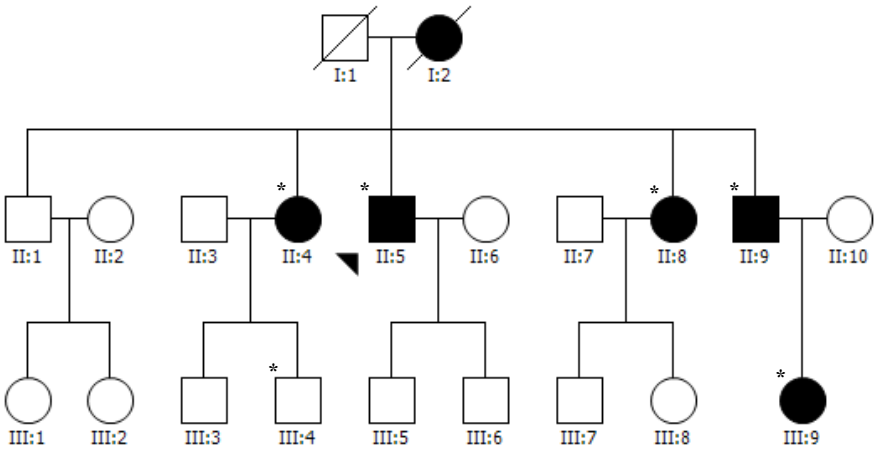


Figure 5.4: Pedigree of family 1 after clinical re-examination. An asterisk top left of a pedigree symbol indicates an individual for which DNA was available.

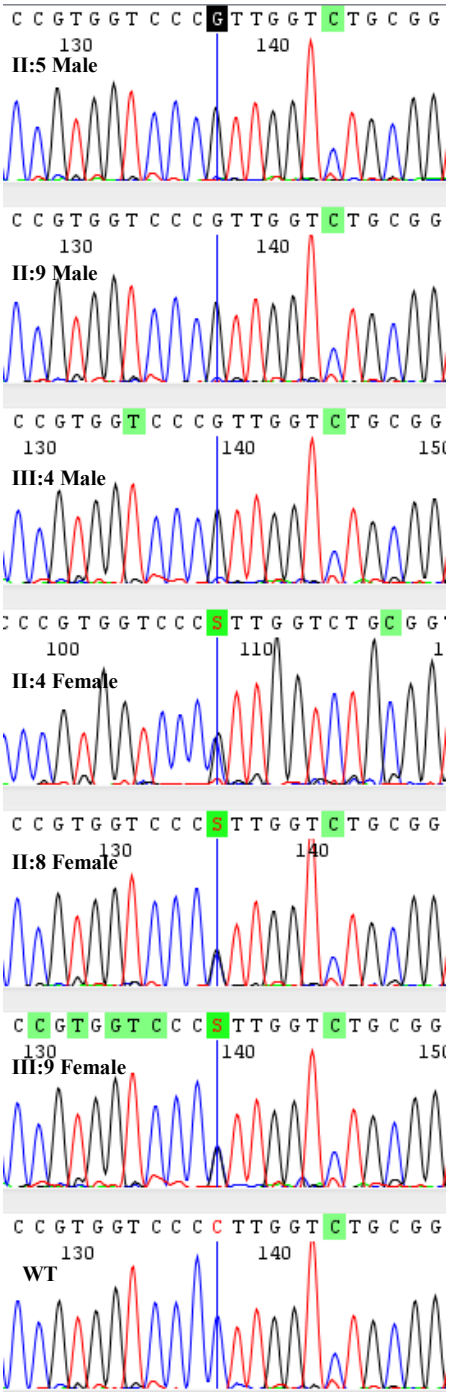


Figure 5.5: Chromatograms of family members showing hemizygous NM_005278: c.651G>C *GPM6B* mutations in affected males and heterozygous variants in affected females.

5.6 Family 2 results

5.6.1 Linkage exclusion analysis by SNPs in family 2

The data obtained from the linkage exclusion analysis excluded the majority of known autosomal dominant SCA genes, with SCA8 and SCA10 not being excluded. SCA8 was already excluded in the family by basis of genetic testing performed at the diagnostic Neurogenetic Laboratory, at that time based at Royal Perth Hospital.

5.6.2 Screening of the SCA10 gene by Genescan

Screening of the *ATXN10* repeat in two affected members of the family by Genescan found that both affected individuals possessed two different-sized normal alleles, eliminating *ATXN10* as a candidate gene.

5.6.3 Analysis of variants identified by exome sequencing in family 2

Exome sequencing allied with linkage analysis and candidate gene screening did not identify any strong candidate disease genes in the family.

5.7 Family 3 results

5.7.1 Linkage exclusion analysis by SNPs in family 3 – autosomal recessive SCA

Linkage exclusion analysis, taking into account the consanguinity of the family excluded all known recessive SCA disease genes except *TTPA*, *PEO1*, and *SETX*. The Friedreich ataxia 2, SCAR2, 3, and 6 loci were not excluded in whole.

A single large deletion was found on chromosome 13 via SNP CNV analysis; however, this region was excluded by linkage analysis.

5.7.2 Exclusion of known ataxia disease genes by exome sequencing

The three remaining known recessive SCA disease genes: *TTPA*, *PEO1 (c10orf2)*, and *SETX* were excluded on the basis of the exome sequencing results, as no variants detectable by NGS were found to be within these genes, which were fully covered to a read depth of >20x. Current literature has only described point mutations, and small deletions and duplications in these genes, all of which are detectable by NGS.

5.7.3 Selection and verification of candidate variants from exome sequencing

Of the variants remaining, three heterozygous variants were detected in the gene ZFHX4, and a homozygous variant NM_018126:c.325A>C (p.T109P) was detected in the gene TMEM33. The variant was found at a 0.01% incidence in the ExAC database, but only in a heterozygous state. The variant within *TMEM33* was not excluded by linkage analysis, and found to segregate with disease status within the available family members

The variants within the gene *ZFHX4* were not excluded by linkage analysis, but did not segregate with disease status in the family, and as such, *ZFHX4* was excluded as a candidate disease gene.

5.7.4 TMEM33

The TMEM33 protein is a predicted transmembrane protein, and is expressed in brain and peripheral nerve. It is a known binding partner of melatonin receptor 1A (MTNR1A), ubiquitin specific peptidase 19 (USP19) and ubiquitin C (UBC),[168-170] and a predicted binding partner of surfeit 4 (SURF4), retention in endoplasmic reticulum 1 (RER1), and pyruvate dehydrogenase (lipoamide) beta (PDHB).[171]

The mutation lies within the second transmembrane alpha-helix domain, and changes a threonine residue to a proline. The insertion of a proline residue into an alpha-helix is a known structure-disruptor. A proline residue has a dihedral angle of a fixed –65° and creates a kink in the protein structure.[172]

5.8 Discussion

5.8.1 Family 1: X-linked HSP

Originally, the inheritance pattern of the disease in this family was diagnosed as autosomal recessive. When all variants identified using exome sequencing were excluded by linkage analysis, other members of the family were re-examined (Lamont PJ, personal communication) and three female members of the family were found to be mildly affected and the pattern of inheritance was re-assigned as X-linked.

A candidate variant in the gene *GPM6B* was subsequently identified. GPM6B has been seen to interact with the protein product of the known disease gene *PLP1*.[167] However, the decision was made that no further functional characterisation would be performed until additional families with similar phenotype and variants within GPM6B were identified.

5.8.2 Family 2: Autosomal dominant SCA

No candidate variants were identified by using exome sequencing coupled with linkage analysis. One possible explanation of this result is that the disease in this family is caused by a novel repeat expansion, which would be undetectable by exome sequencing.

DNA samples from affected individuals from the family have been sent to Dr Daniel MacArthur at the Broad Institute, Boston, for whole-genome sequencing using a specialised technique that Dr MacArthur is developing to identify triplet repeat expansions from next generation sequencing data. The sequencing method to be employed in this further research is thus, a novel, experimental technique.

5.8.3 Family 3: Autosomal recessive SCA

A mutation NM_018126:c.325A>C (p.T109P) in a candidate disease gene *TMEM33* has been identified by combining linkage analysis with exome sequencing in this family.

The mutation inserts a proline residue into a transmembrane alpha helix. Insertions of proline residues into alpha helices can cause recessive disease, as seen in some desminopathy patients.[173] The protein product is of unknown function, but has been shown to interact with the MTNR1A, SURF4, RER1 and PDHB proteins. Of these interacting proteins, SURF4 is expressed in the both endoplasmic reticulum (ER) and the Golgi apparatus, and interacts with ER-Golgi intermediate compartment proteins. RER1 is expressed in the Golgi apparatus, and is involved in the retention of ER membrane proteins in the ER and retrieval of ER membrane proteins from the early Golgi compartment to facilitate gamma-secretase complex assembly.

Further investigations of the *TMEM33* gene have been halted until another family with a similar phenotype has been identified. This family resides in Indonesia, and as such, additional DNA for further analysis is hard to come by. However, the family has been contacted for follow-up studies. additional exomes are planned to be sequenced and the family will be re-analysed as a trio, before applying the affecteds-only linkage filter from section 5.6.1.

5.8.4 Applications of NGS – difficulties in identifying novel disease genes

In this chapter, three families were investigated with a combination of next generation and traditional disease genes discovery methods. However, no candidate gene could be identified for the dominant SCA family and although candidate genes were identified in the recessive SCA family and the X-linked HSP family, further work was suspended.

Recent data from the FORGE Canada consortium rare-disease gene-discovery project shows that out of the 19 autosomal dominant disease families studied, an aggregate success rate of 37% was achieved, out of which only a single novel disease gene was identified.[142] As such, it was disappointing, but not unexpected that a candidate disease gene was not identified in the autosomal dominant SCA family that I investigated.

Additionally, the FORGE Canada consortium achieved a 70% success rate when investigating their 60 consanguineous families, and a 45% success rate when investigating their 62 non-consanguineous families. More importantly, the highest success rate achieved (94%) was when sequencing multiple unrelated individuals or families affected by the same very rare, but highly recognizable clinical condition[142] – a resource that was unavailable to me when I was investigating families 2 and 3.

To this end, the phenotype description of affected members of family 3 have been uploaded to the Australasian Neuromuscular Network's call for patients portal, in an effort to identify additional unrelated individuals with the same condition. A recent publication has also shown that sequencing of trios (mother, father, affected individual) have an increased success rate as compared to just sequencing of a single affected individual.[174] This will be performed in the X-linked HSP family to confirm *GPM6B* as a candidate disease gene.

Chapter 6

Development and validation of a high-throughput neurogenetic sub-exomic sequencing capture panel for research and diagnostic applications

6.1 Summary

In this chapter I describe the construction, validation and deployment of a 336-gene cardiac and neurogenetic disease targeted capture and sequencing panel. A global success rate of 29.6% was achieved, in line with other targeted capture cohort studies. This success rate rose to 34% when only considering the set of patients screened on a purely diagnostic basis. This targeted capture panel is now deployed as a front-line test at the Neurogenetics Unit, Department of Diagnostic Genomics, PathWest Laboratory Medicine, Department of Health, Western Australia.

6.2 Introduction

6.2.1 Neurogenetic disorders and challenges to diagnosis

The neurogenetic disorders are a genotypically and phenotypically broad spectrum of diseases affecting the nervous system and skeletal muscle.[175; 176]

There are three major challenges for molecular diagnosis of these disorders using the mainstay of molecular diagnostic laboratories; Sanger sequencing. Firstly there is the high level of genetic heterogeneity. For example, more than 60 genes have been associated with inherited peripheral nerve disease,[147] and more than 40 with muscular dystrophy.[177]

Secondly, many of the largest human genes, for example, dystrophin (*DMD*), nebulin (*NEB*), the ryanodine receptor (*RYR1*) and especially titin (*TTN*) are mutated in these conditions.[178] This leads, in the diagnostic setting of limited budgets, to an insurmountable burden of laboratory work needed to screen these genes using Sanger sequencing.[178; 179]

Thirdly, there is the challenge of multiple clinical phenotypes, or different disease inheritance patterns being associated with mutations in one gene. An extreme example is the lamin A/C gene (*LMNA*) (OMIM #150330), which has been associated with 12 different disease phenotypes, including both neurogenetic and non-neurogenetic diseases, and both autosomal dominant and autosomal recessive inheritance.[180]

These three factors can lead to difficulty selecting the correct genes to sequence for a given patient. The molecular cause of the disease in the patient may not be identified because the wrong gene/s are examined and because there is an insufficient diagnostic budget available to have all the possible candidate disease genes sequenced with Sanger sequencing.

Thus, it is extremely difficult to obtain a genetic diagnosis for a large percentage of patients. Current best practice in many diagnostic labs worldwide was to only screen the most common known disease genes for mutations using Sanger sequencing. Even then, in the case of large genes e.g. *RYR1*, only a few exons known to be mutation hotspots might be screened. When this thesis was started, this approach was the only accredited cost-effective way of achieving a genetic diagnosis for patients, but leaves much to be desired in the efficacy of diagnosing the majority of cases in such a genetically heterogeneous group of diseases.

As an example, this paucity of genetic diagnoses can be seen in the cohort of patients at the Royal Perth Hospital Neurogenetic Laboratory, where there are over 900 patients diagnosed with nemaline myopathy still without a genetic diagnosis, and over 400 patients diagnosed with Charcot-Marie-Tooth disease without a genetic diagnosis (Personal communication, Laing, NG).

In contrast with traditional Sanger sequencing, next generation sequencing technologies efficiently sequence multiple human disease genes simultaneously. This may be by sequencing the genome,[181] the exome[182] or a sub-exomic panel of targeted disease genes.[183] Genome sequencing is currently still too expensive, and generates too much data for analysis per patient for use in routine diagnostic laboratories. Exome sequencing sequences most of the coding region of all human genes, allows for genotype phenotype diversity and correction of misdiagnoses[184] as well as iterative data mining as new disease genes are identified.

However, exome sequencing, like genome sequencing gives reduced throughput of samples on available hardware at greater expense per sample and generates more data to analyse than targeted gene panels.[183] On the other hand, a very restricted panel of genes targeted for one clinical phenotype does not allow for diversity of genotype-phenotype associations or correction of misdiagnosis.[185]

Hence, a neurogenetic sub-exomic capture panel targeting the known neurogenetic disease genes was designed to decrease the cost and time of sequencing, and provide a more specific dataset for quicker analysis. A published example is the 267 neuromuscular disease gene panel tested by Vasli *et al* on eight positive controls and eight prospective samples,[186] which offers a middle path of relatively high throughput, while allowing for genotype-phenotype diversity and correction of misdiagnoses via "diagnosis by sequencing".[184]

To this end, one portion of the work in this thesis was devoted to the design and testing of a "neurogenetic sub-exomic sequencing" (NSES) array, to use in diagnostic applications, the NGS bioinformatic pipeline developed in Chapter three. All the known

neurogenetic disease genes up to a certain date, December 2012, were collated, and probes designed to capture the exonic sequence for subsequent next-generation sequencing those genes where known mutations would be detected by NGS.

6.2.2 Aims

1) To design a capture array targeted against all the currently known neurogenetic disease genes in which known mutations could be detected by next generation sequencing.

2) To test the efficiency of multiplexing patient sample captures for the purposes of reducing diagnostic cost.

3) To increase the efficiency of diagnosis by increasing the number of genes screened

4) To translate next generation sequencing into routine clinical application.

6.3 Materials and Methods

Written informed consent was obtained for participation in this study, which was approved by the Human Research Ethics Committee of the University of Western Australia.

Control samples were selected from the cohort at the Neurogenetics Unit, Department of Diagnostic Genomics, PathWest Laboratory Medicine, Department of Health Western Australia. Twenty patients with known small-scale variants (variants involving single nucleotides or only a small number of nucleotides and small indels) and 12 patients with known copy number variants were sequenced on the NSES panel.

A further 450 patients without a genetic diagnosis were selected to be screened on this panel. Of those patients: 308 were screened on a diagnostic basis, with 40 having had

previous genetic testing. 142 patients were screened on a research basis, with 46 having had previous genetic testing.

6.3.1 Capture panel design

The majority of neurological disease genes (n=254) chosen to be targeted were based on the December 2012 freeze of the Neurogenetic Disorders gene list[177], excluding disease genes with mutations not detectable by next generation sequencing, for example, repeat expansions (myotonic dystrophies, spinocerebellar ataxias., etc), repeat deletions (facioscapulohumeral muscular dystrophy) or deletions and duplications in highly homologous genes (the SMN1 spinal muscular atrophy locus). In addition to this initial large, base disease gene set, known disease genes in the foetal akinesia spectrum were added, as well as a set of novel/unpublished and candidate disease genes. This resulted in a total of 277 neurological disease genes. 59 cardiomyopathy disease genes were added since there is considerable genetic overlap between neurogenetic and cardiomyopathy disease genes, for example, the myosin heavy chain 7 gene *MYH7*[187] and *TTN*[188; 189] and therefore, many of the same genes would have had to be repeated on smaller panels targeted solely to either neurological or cardiomyopathic disorders. Therefore a total of 336 genes were targeted. The genes targeted in this panel can be found in Appendix B, arranged in alphabetical order according to disease category. Duplicate entries are present when a disease gene is associated with multiple disease phenotypes.

The 336 genes targeted contained 6772 exons, including non-coding exons and 5' and 3' untranslated regions. Exon start and end positions with 50 base pairs flanking intronic sequence of the disease gene set were extracted from the RefGene database of the hg19 build of the UCSC human genome. These positions were sent to Life

Technologies where the TargetSeq™ liquid phase capture probes were designed and manufactured. The total length of sequence targeted for capture was 2.8Mbp.

6.3.2 Capture and Sequencing on the Life Technologies Proton

All samples were sequenced using an Ion Torrent Proton semiconductor sequencer (Life Technologies) at the Lotteries West State Biomedical Facility, Genomics node (LSBFG), located at Royal Perth Hospital, Perth, Western Australia.

Pre-capture libraries were prepared using NEB NeXT Ultra reagents with local modifications (New England Biosciences, (NEB)) and Ion Xpress Barcode Adapters with T-overhangs. After adapter ligation, libraries were amplified for 6 cycles using the Q5 polymerase (NEB) and quantified on a Bioanalyser 2100 (Agilent).

Samples were captured as per the Ion TargetSeq Custom Enrichment Kit (Life Technologies) standard protocol with minor modifications. Equal amounts of 16 libraries were pooled to give a total of 500ng for each capture. The pooled libraries were dried down with blockers then combined with capture probes and hybridised for 64 to 72 hours at 47^{o}C, followed by washing three times. Post-capture libraries were subject to 10 cycles of amplification using Q5 polymerase (NEB) and size-selected at approximately 275bp using 2% E-gels (Invitrogen). A final library concentration of 7.5pM was used for the templating reaction.

Templated Ion Spheres were made using an Ion P1 200 V2 or V3 template kit on a OneTouch2 (Life Technologies). The templated beads were sequenced on a P1v2 sequencing chip (Life Technologies) using an Ion P1 200V2 or V3 sequencing kit for 520 flows.

6.3.3 Variant calling and annotation

Base calling, mapping and variant calling were performed using Torrent Suite 3.6.2 or 4.2 against the build GRCh37 human reference genome. Variant calling was performed using the Torrent Variant Caller (TVC) with preset high stringency settings.

Variants from the Ion Proton sequencer were analysed in the Centre for Medical Research, University of Western Australia, Harry Perkins Institute of Medical Research using the ANNOVAR[128] based pipeline described in chapter 3.4.3. Variants within the dbSNP137Common variant database (containing variants present in the database at a frequency of one percent or greater) as well as variants found in an in-house common variant list were excluded. The remaining variants were first annotated against the EncodeGencodeBasicV14 gene annotation set. Minor allele frequency annotations were against the Thousand Genomes Project[136] and NHLBI Exome Sequencing Project (ESP) Exome Variant Server[26] frequencies, if the variant was present in these databases. Variants were further annotated with SIFT, PolyPhen and MutationTaster variant pathogenicity predictions. This sequencing pipeline was the one described in Chapter 3.

After filtering and annotation, variants were compared against the HGMD Professional database (http://www.hgmd.cf.ac.uk/ac/index.php) to identify any known mutations. Potential splicing effects were assessed using the Alamut software (Interactive Biosoftware).

First-pass diagnostic samples were also screened at the Department of Diagnostic Genomics using BenchLab NGS (Cartagenia), which allowed variant filtering, with the addition of an automatically updated common variant database, streamlined known disease variant identification and variant prioritisation based on disease phenotype.

Variant pathogenicity was ranked on a defined set of criteria. Highest ranking was given to variants described as disease-causing in previously published literature and matching the disease phenotype and inheritance, followed by novel variants within known disease genes associated with the phenotype and then known disease variants in other genes not previously associated with that disease phenotype and finally novel variants in disease genes not known to be associated with the disease phenotype.

6.3.4 Coverage analysis

Analysis of target coverage was done using the BedTools[190] software suite. Mapped sequence read data in the BAM files were processed to identify the total number of reads mapping to target regions. This was done to identify regions of genes where coverage was absent, or insufficient to the point where variants could not be reliably called.

This was performed in conjunction with analysis of coverage using IGV to confirm coverage across exons within the targeted genes.

6.3.5 Copy number variation analysis

Detection of copy number variants (CNVs) from the targeted panel sequencing data was performed using the unpaired analysis pipeline within the FishingCNV v2.1 software suite[138], against sets of sex-specific controls (n=100 for females, n=100 for males) if an X-linked disease was suspected, or against a normalised set of 200 controls.

CNV results were examined in both the FishingCNV output and Integrated Genomics Viewer[139] (IGV) for samples where there were either single heterozygous variants in genes known to cause recessive disease, or in patients where no likely candidate small-scale variants were found in associated disease genes. In the case of these patients, examination of CNV results was limited to *DMD* and *PMP22* if the phenotype was compatible, plus additional genes on a per-patient basis.

6.3.6 Sanger validation

Candidate disease-associated variants generated from the next generation sequencing of the targeted panel of genes were sequenced by Sanger sequencing by the Diagnostic Genomics Laboratory, PathWest, Western Australia to confirm the presence of the variant in the patient's sample, or demonstrate that the next generation sequencing result was invalid.

6.4 Results

6.4.1 Sequencing statistics

Samples from a total of 482 patients were sequenced using the targeted 336-gene panel: 32 positive controls (20 with known small-scale variants and 12 with known copy number variants) and 450 unrelated patients screened prospectively.

Samples from 16 patients could be sequenced simultaneously in one Ion Proton sequencing run using the Ion P1 V2 chemistry. A very high mapping rate (>98% of reads mapped to the human genome) was obtained, with approximately 80-85% of reads mapping to the defined target regions. This is consistent with previous experience using TargetSeq probes. The average sequencing depth per-sample was 217-fold, with more than 92% of the target regions covered to 20-fold or greater depth.

This compares favourably to our previous experience of exome sequencing using the TargetSeq exome capture platform on a SOLiD 5500XL sequencer in which an average sequencing depth of 66-fold per set of 12 samples analysed and 84% of targets covered to 20-fold or greater depth was obtained.[107; 108; 191] The increase in coverage using the sub-exomic targeted panel led to many exons that were not well covered by exome sequencing being covered to a depth of greater than 20-fold in the targeted panel sequencing.

A shift to the Ion Proton V3 chemistry after sample 368 resulted in an average sequencing depth-per-sample of 309-fold, with more than 96% of the target regions covered to 20-fold or greater depth.

6.4.2 Positive control hit rate

A 90.0% hit rate for small-scale variant positive controls was achieved with the NSES array (Table 6.1). Of the two controls where the correct result was not found, one had the mutation in an exon of a gene that was known to not capture well, and the second mutation was in a homopolymer region, which are known to not sequence well on the Proton™ platform. However, the homopolymer mutation was later identified in a different, diagnostic sample sequenced using the improved V3 Ion Proton chemistry.

A 91.6% hit rate for the CNV positive controls was achieved using the FishingCNV software suite (Table 6.1). For the single control with a false result, a deletion of exons 2-4 of *PAFAH1B1* was detected when the correct result was a deletion of exon 3 alone. Notably, all 5 *PMP22* gene deletion/duplication controls screened were correctly identified.

6.4.3 Prospective sample hit rate

6.4.3.1 Definitive mutations

Of the 450 patients screened prospectively, a global success rate of 29.6% was achieved.

86 patients had previously been screened using the standard Neurogenetics Unit Sanger sequencing based protocol for disease genes associated with their phenotype (up to 8 disease genes) with no mutations identified, and 364 had had no previous analysis. Mutations were identified in 19 (22.1%), of the 86 patients that had previously been screened with the routine laboratory protocol and 114 (31.3%) of the 364 patients that had had no previous analysis.

If results are broken down into samples screened on a diagnostic or a research basis, 308 patients were screened on a diagnostic basis, and 142 were screened on a research basis. In these two categories, mutations were identified in 34.1% of diagnostic samples, and 19.7% of the research samples. (Table 6.2)

Mutations were identified in 71 genes in the patient cohort – 21% of the genes on the panel. Of these 71 genes, 48 were not previously analysed or only partially analysed in the diagnostic laboratory.

Of the six largest disease classes screened, a successful diagnosis was achieved in 29% (21/73) of myopathy patients, 42% (22/52) of muscular dystrophy patients, 26% (12/47) of peripheral nerve disease patients, 40% (16/40) of the hereditary spastic paraplegia patients, 31% (10/32) of the cardiomyopathy patients and 50% (11/22) of the channelopathy patients. A single PMP22 duplication was identified in a prospective case that had had no previous screening. (Table 6.2)

The clinical diagnoses of the patients for whom mutations were identified were wide-ranging. For muscle diseases, the diagnoses ranged from congenital myopathy to recurrent rhabdomyolysis, from foetal akinesia to adult-onset limb girdle muscular dystrophy. For the central/peripheral nervous system disorders, the diagnoses ranged from Charcot-Marie-Tooth disease to non-5q spinal muscular atrophy.

Table 6.1: Table of positive control hit-rate.

	Number	Detected	Undetected	Hit rate
Small-scale variant (SSV) Positive control	20	18	2	90.0%
CNV positive control	12	11	1	91.6%

Table 6.2: Summary table of molecular diagnoses achieved per disease class, grouped by clinical diagnosis.

	Diagnostic			Research		
Disease category	**Total referred**	**Total solved**	**%age solved**	**Total Referred**	**Total solved**	**%age solved**
Anterior horn cell disease	10	2	20%	6	2	33%
Arthrogryposis	11	4	36%	17	2	12%
Cardiomyopathy	32	10	31%	-	-	-
Channelopathies	22	11	50%	-	-	-
Hereditary spastic paraplegia	40	16	40%	-	-	-
Metabolic myopathy	4	0	-	-	-	-
Movement disorder	1	0	-	-	-	-
Muscular dystrophy	52	22	42%	12	6	50%
Myopathy	73	21	29%	39	5	13%
Neuromuscular junction	4	2	50%	3	0	-
Peripheral nerve disease	47	12	26%	24	5	21%
Rhabdomyolysis	-	-	-	29	5	17%
Nonspecific	12	5	42%	12	2	17%
Totals	308	105	34%	142	28	20%

Anterior horn cell disease includes: distal neuronopathy, motor neuron disease, spinal muscular atrophy
Arthrogryposis includes: arthrogryposis, distal arthrogryposis, foetal akinesia deformation sequence (FADS)
Cardiomyopathy includes: arrhythmogenic right ventricular dysplasia, dilated cardiomyopathy, hypertrophic cardiomyopathy, long QT
Channelopathies includes: malignant hyperthermia, hypokalemic periodic paralysis, myotonia/paramyotonia congenita, ataxia, episodic ataxia type 2
Metabolic myopathy includes: glycogen storage disease, mitochondrial myopathy
Movement disorder includes: dystonia
Muscular dystrophy includes: congenital muscular dystrophy, dystrophinopathies, Emery-Dreifuss muscular dystrophy, non-4q facioscapulohumeral muscular dystrophy, limb girdle muscular dystrophy, muscular dystrophy, rigid spine muscular dystrophy,
Myopathy includes: Bethlem myopathy, central core disease, congenital myopathy, distal myopathy, multicore-minicore myopathy, myofibrillar myopathy, myopathy, nemaline myopathy
Neuromuscular junction includes: congenital myaesthenic syndrome
Peripheral nerve disease includes: Charcot-Marie-Tooth disease, congenital hypomyelination, neuropathy
Nonspecific includes: ataxia, cramps/myalgia, malignant hyperthermia, very long chain acyl-CoA deficiency

6.4.4 Coverage statistics

Like all capture methods targeting exons, deep intronic regions are not targeted for capture as a feature of design, and as such we will not be able to see potential disease-causing variants deep within introns. A total of 162 exons out of 6772 (2.41%) need to be Sanger sequenced to ensure complete coverage of the known disease genes included on the panel.

These exons include those where capture is inefficient or incomplete, as seen in Figure 6.1 for *FKRP*. This is a known limitation of probe-based capture technologies, where capture efficiencies decrease at extremes of GC content.[192]

6.4.5 Additional sequencing findings

Homopolymer repeat regions within the genome are not sequenced with good efficiency. In particular, regions of three or more guanine or cytosine nucleotides in a row tend to be called with deletions of one or more nucleotides. These sequencing errors are ignored or filtered out in the final analysis.

In addition, the *NEB* and *TTN* genes contain repetitive exon clusters that probes were not designed for or were designed for but did not map. This result can be seen in Figure 6.2, where although probes were designed for a few exons within the repetitive exon cluster in the *NEB* gene, reads either are not mapped due to quality thresholds, or are mapped non-uniquely and as such are not considered by the variant calling engine. Also for this reason, the SMN1/2 gene was not included in the capture set.

6.4.6 Additional diagnostic findings

Notably, the panel identified causative mutations in four patients, where the gene containing the causative mutation had already been Sanger sequenced in a diagnostic laboratory and the mutations had been missed.

In addition, the Sanger fill-in approach was validated for limb-girdle muscular dystrophy patients without a diagnosis from using the panel. In 2 of 13 limb-girdle muscular dystrophy probands, *FKRP* mutations were identified. Another successful application of the Sanger fill-in approach was in a proband with a single recessive *B3GALNT2* mutation. Exon one, which was known to be badly covered, was screened by Sanger sequencing, and a second recessive mutation was identified.

In another case a single recessive *RYR1* mutation was identified. Subsequent cDNA analysis identified that the other *RYR1* allele was a null allele, not expressed for reasons unknown.

Figure 6.1: IGV screenshot of the single coding exon of *FKRP*, showing capture inefficiencies due to the high G/C content of the exon. This can be seen by the lack of reads mapping (light red and blue boxes) and subsequent gaps in coverage (gaps in grey graphs). Such exons may be Sanger sequenced to ensure complete coverage of known disease genes.

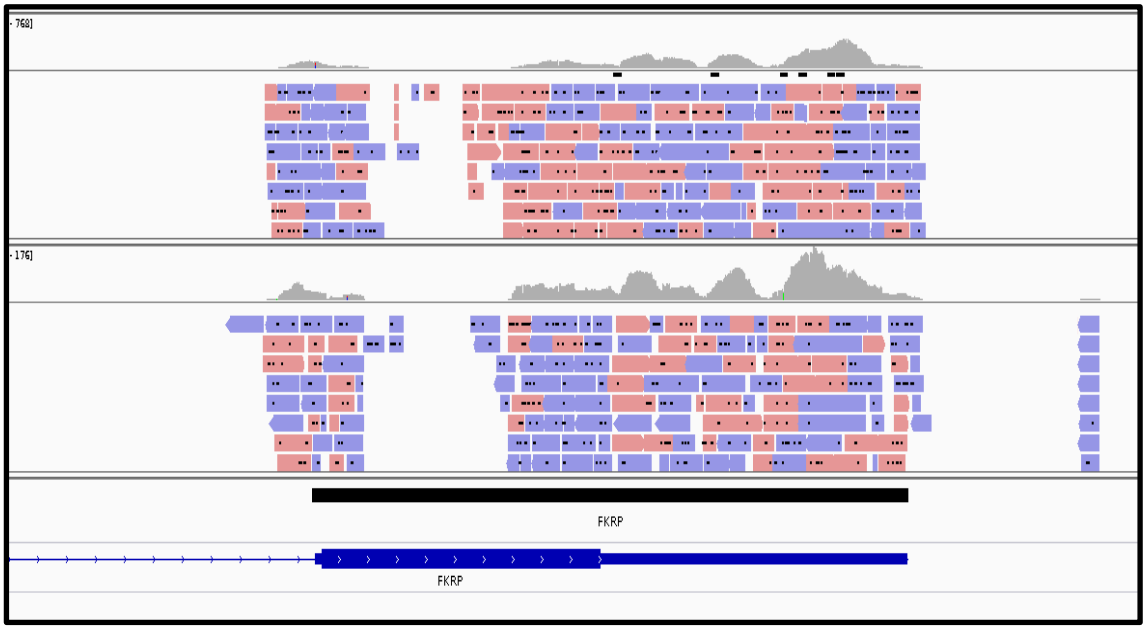
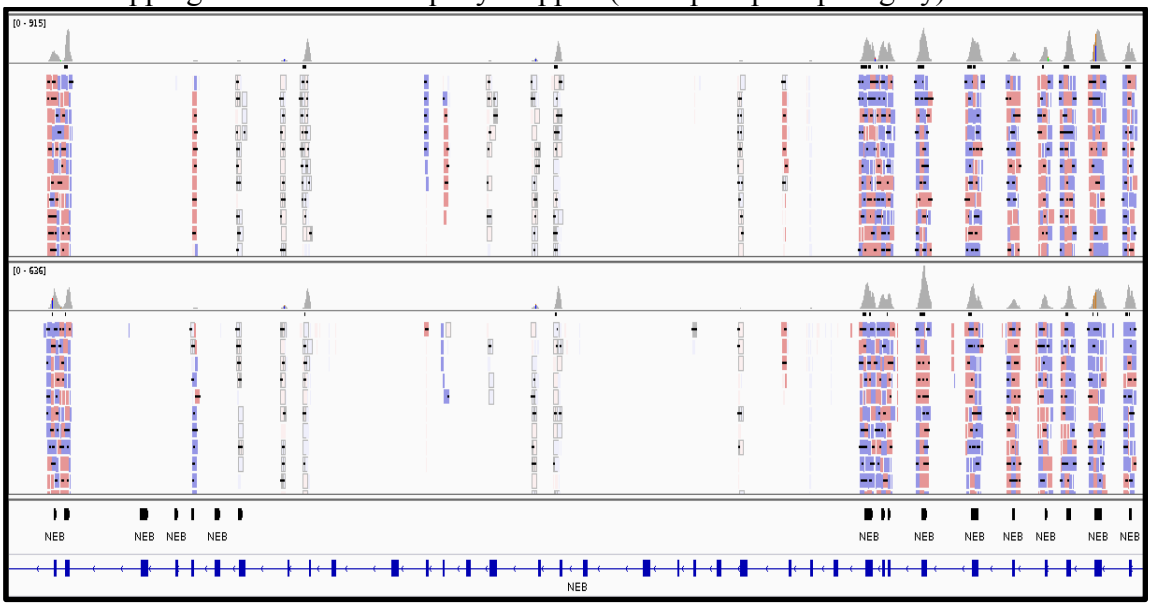


Figure 6.2: IGV screenshot of the repetitive exon cluster in the *NEB* gene, showing where probes (row in black) were not designed over the repetitive exons (in dark blue). Reads mapping here are non-uniquely mapped (clear/pale pink/pale grey).



6.5 Discussion

6.5.1 Benefits of using a gene panel containing a large number of known disease genes

We developed a next generation sequencing gene panel targeting 336 known disease genes across a large spectrum of genetic neurological disorders and cardiomyopathies. We chose this approach for a number of reasons; including providing a relatively high throughput, taking account of genotype-phenotype diversity[183] and allowing "diagnosis by sequencing.[184] Our results demonstrate that the large multi-disease targeted panel is capable of screening the majority of known genes for multiple groups of neurological diseases in parallel.

Causative mutations were identified in 133 of the 450 prospective samples, equivalent to a success rate of 29.6%. When split into diagnostic and research samples, mutations were identified in 34.1% of diagnostic samples, and 19.7% of research samples.

Of the 450 prospective samples screened, 86 had undergone routine screening of known disease genes with negative results. Mutations were identified in 22.1% of these samples whereas mutations were identified in 31.3% of the unscreened prospective samples. The difference in success rates between the two classes of prospective samples could be explained by the identification of a large number of mutations in genes that would have been routinely screened in these patients, for example, *ACTA1*.

The large multi-disease panel helps overcome situations of diagnostic difficulty. Salient examples of this from the cohort we studied include one patient diagnosed with severe spinal muscular atrophy who was shown to have a mutation in the skeletal muscle actin gene (*ACTA1*) (a situation which has been previously described);[9] two adult patients

diagnosed with limb girdle muscular dystrophy who were shown to have known mutations in *ACTA1*, a gene that would not usually be considered a candidate for the phenotype. These "diagnosis by sequencing" results highlight the power of unbiased sequencing of multiple disease genes across many classes of neurological disorders, as opposed to screening a limited subset of disease genes based on the clinical diagnosis.[184] Diagnosis by sequencing may allow the patient to avoid expensive and invasive diagnostics such as muscle biopsy.

The panel also found mutations in genes already screened by diagnostic laboratories, in four probands. The inclusion of commonly screened genes on the panel allows a double-checking of previous Sanger sequencing based results (if any) at effectively no additional cost. This allows for error correction in these rare cases. Possible reasons for Sanger sequencing not identifying causative mutations are allele dropout (where one or more alleles are not present) due, for example, to SNPs underneath the chosen PCR primers,[193-195] or mutations that are missed due to simple human error.

6.5.2 Cost effectiveness

Cost-effectiveness is a major point in favour of next generation sequencing technologies. Of the 71 genes in which mutations were found, 48, or more than two thirds, have not been part of routine screening by the diagnostic Laboratory because of labour and consumables costs. The ability of next generation sequencing diagnostics to identify mutations in disease genes not previously able to be screened is a major step in cost-effectiveness. Neveling *et al.* estimated that once more than three genes had been screened by Sanger sequencing with a negative result, exome sequencing is the more cost-effective choice.[196] Similar arguments have been proposed by others.[197; 198]

Having targeted gene panels made is expensive; therefore making one gene panel targeting a large set of genes is more cost-effective. One large panel also allows the use of one test for most samples analysed by the diagnostic laboratory, simplifying workflows.[183; 186] The reduced amount of DNA targeted for sequencing in a targeted panel compared to an exome (2.8Mbp versus ~50Mbp in an exome) will always allow higher sample throughput and sample coverage than whole exome or whole genome sequencing. Using the current panel, 16 samples can be screened to more than double the average coverage of exome sequencing, and more than three times the average depth of coverage with the now currently employed V3 Ion Proton sequencing chemistry. As sequencing costs lower and efficiency rises, even higher throughput can be realised without lowering the sequencing depth.

6.5.3 Reasons for the success rate in identifying mutations of only around 34% in diagnostic samples

The 34% success rate we obtained in diagnostically screened samples is in line with the results of other mixed cohorts,[199; 200] but falls short of the up to 80% success rate when a highly selected patient cohort from a single class of disease was examined.[201]

One undoubted reason for the success rate not being higher, is that many disease genes for genetic neurological disorders were not known when the panel was designed, and therefore could not be included. Our estimate is that more than 100 relevant novel disease genes, that would now be included, have been discovered since the design of the panel, with others remaining to be found.

Some patients may miss out on a diagnosis because of the lack of coverage in ~2.5% of the targeted exons. These regions include high GC-content regions and highly

homologous or repetitive regions of the genome that are not captured well by current technologies.[202] These "missing" exons may be filled in by Sanger sequencing. It is anticipated that of these "missing" exons, the whole set does not need to be screened by Sanger sequencing in every patient, instead, a select number of exons in disease genes known to be associated with the patient's phenotype would be screened.

Another way to fill in the missing exons might be by a targeted panel approach using a different technology such as Haloplex PCR.[203]

The 'missing exons' problem may be partially overcome by improved sequencing methods. The shift to the Life Technologies Ion Proton V3 sequencing chemistry shows increased coverage; from an average of 217-fold to 309-fold, and the percentage of targets covered to 20-fold or greater increased from 92% to 96%. This increase in coverage means that many exons that were not previously covered well are now covered adequately. Two companion NGS methods are being investigated for this 'gap filling' work; increasing the concentration of capture probes in exons that are not covered well, and a separate capture set targeting only those exons.

Like all capture methods targeting exons, deep intronic mutations will also be missed, as may be mutations at homopolymer repeat regions, such as the missed positive control.

Synonymous variants are also problematic and not confirmed by Sanger sequencing unless they are known to be disease-causing, due to the difficulty of determining the pathogenicity of synonymous variants.[204]

Lastly, identification of disease-causing copy number variations remains problematic. The inclusion of CNV samples in the positive control set was a proof-of-concept exercise, as there was little published literature at the time on calling copy number variants from targeted capture and sequencing.[186; 205; 206] The success in calling the CMT1A/HNPP duplication/deletion of *PMP22* in 100% of the positive controls, plus a *PMP22* duplication in a single prospective patient was encouraging.

In the positive controls diagnosed with DMD/BMD, duplications down to 8 exons in size, and deletions down to a single exon were detected. However, currently, detection of copy number variation is insufficiently sensitive and too labour-intensive for routine application in a diagnostic setting. Therefore, copy number variant data generated using the current pipeline should be used only as a "value-add" proposition; for example for patients where one known or likely recessive mutation has been identified as opposed to a front-line analysis for all samples.

6.5.4 Future trends in gene panel sequencing molecular diagnostics for genetic neurological disorders

The future of next generation sequencing diagnostics for neurogenetic diseases is fluid.

A large multi-gene multi-disease targeted panel, such as used here, offers high-throughput, low-cost, accurate identification of variants in known disease genes such that use of such a panel may be argued to be ethically no different to screening each of the genes one at a time by Sanger sequencing. It minimises the chances of incidental findings by only screening genes that are relevant to a subset of disorders. Further bioinformatic filters applied at the analysis stage can limit the number of variants

displayed to a smaller subset e.g. genes implicated in familial spastic paraplegia, or nemaline myopathy.

In 2013, the American College of Medical Genetics and Genomics released a set of guidelines for the reporting of incidental findings in genomic tests.[12] Applicable to this panel are guidelines for diseases where preventative measures and/or treatments were available. It also includes disorders in which individuals with pathogenic mutations might be asymptomatic, and the reporting of variants that have been previously identified as pathogenic. The cardiomyopathies and malignant hyperthermia susceptibility are prime examples on this panel.

However, the large amount of phenotypic heterogeneity in the neurogenetic disorders may preclude aggressive bioinformatic filtering. If too aggressive filtering is applied, results such as the single *ACTA1* mutant in a supposed spinal muscular atrophy case would have been filtered out and ignored. A more balanced approach may be appropriate when dealing with known disease gene panels, as every gene screened is a known disease gene.

A current problem for targeted panels such as described here, as opposed to whole exome sequencing or whole genome sequencing, is the need to update the panel as novel disease genes are identified. It can however be anticipated that the identification of disease genes will asymptote towards completion and in time the need to reiterate the panel will become less frequent.

Good clinical information is essential to obtain the best outcomes from next generation sequencing diagnostics. Some mutations such as repeat expansions or contractions or

duplications or deletions, which are common causes of genetic neurological disorders, are only diagnosable at present by non-next generation sequencing methods and this is likely to remain the case for some time. Therefore accurate clinical diagnosis is needed to decide which patients are likely to have mutations detectable by next generation sequencing technologies and which are not and determine which patients should have next generation sequencing. Vasli *et al.* also argue that accurate clinical diagnosis is invaluable to help decide which of the large number of candidate variants identified for each patient is likely to be the disease-causing mutation.[186] Unless the clinical information is accurate, a variant dismissed as not causing the disease in the patient, could be the causative mutation. This therefore requires close consultation between the diagnostic laboratory and the referring clinician.

In conclusion, the overall result of using the 336-gene panel is that answers can be obtained for patients who had none from previous routine screening, or who would have had none from the Neurogenetic Unit's routine screening protocols. Nevertheless, the use of the targeted panel is not a panacea. It obtains accurate diagnosis for a higher percentage of patients than was previously possible, but still many patients remain without a diagnosis. In the future, further improved knowledge of disease genes and technology may also provide answers for many of these patients.

Chapter 7

Titin mutations identified as a cause of recessive minicore disease

7.1 Summary

In this chapter, in collaboration with UK and Australian research labs, I describe the application of next generation sequencing to the discovery of mutations within the gene encoding the giant sarcomeric protein titin that cause a novel phenotype of minicore myopathy. The identification of these mutations expands the existing phenotypes of titin disease and minicore myopathy.

7.2 Introduction

Minicore myopathy is a congenital myopathy of early onset, defined pathologically by the presence of multiple areas of reduced mitochondrial oxidative activity in muscle biopsies. There are currently three types of minicore disease defined in the Online Mendelian Inheritance in Man (OMIM) database and one additional type in the literature, for a total of four. These are:

(1) Minicore myopathy with external ophthalmoplegia (OMIM #255320), caused by mutations in the ryanodine receptor 1 (*RYR1*) gene.

(2) Myopathy, congenital, with fiber-type disproportion (OMIM #255310) which can also display minicores in pathology, caused by mutations in the selenoprotein 1 (*SEPN1*) gene.

(3) Multi-minicore Disease (MmD) with variable cardiac involvement, caused by mutations within the myosin heavy chain 7 (*MYH7*) gene.[207]

(4) Minicore myopathy, antenatal onset, with arthrogryposis, with no disease gene associated (OMIM 607552).

Minicore disease caused by mutations in the *RYR1* gene is mainly inherited in a recessive manner, but dominantly inherited mutations have also been observed.[208] *SEPN1* mutations associated with a minicore phenotype are inherited recessively and

most mutations are truncating, with only a few missense mutations affecting essential domains of the protein.[209; 210]

Related diseases are central core disease (CCD) and core-rod disease (CRD). These two diseases share the pathological feature of cores in muscle biopsies. Central core disease is currently known to be caused by mutations within the skeletal muscle actin gene (*ACTA1*)[42] and *RYR1*.[211] Core-rod disease is currently known to be caused by mutations in the *ACTA1*[42], *CFL2*,[88] *KBTBD13*[95], *NEB*[212] and *RYR1*[85] genes.

7.2.1 Titin

Titin (TTN) is a giant (some isoforms are almost 4000kDa) sarcomeric protein that among its many functions gives striated muscle its ability to return to its resting length by passive forces.[213; 214] It spans a half-sarcomere, reaching from the Z-disk, within which the N-terminus of titin is anchored, to the M-line. The titin protein has been divided into four separate regions, the Z-disc, I-band, A-band and M-line regions, based on position within the muscle sarcomere. Circled in Figure 7.1 are the Z1Z2 domains, anchored to the telethonin (TCAP) protein within the Z-disk.[215] The part of the titin protein immediately C-terminal of the Z-disk is bound to the thin filament, with the rest of the I-band functioning as a molecular spring.[213; 214; 216] The A-band region of titin is known to bind myosin,[217] but has been reported to have many functions beyond that, from regulation of sarcomere length as a "molecular ruler"[218] to regulation of thick filament structure.[219]

The TTN protein is composed primarily of approximately 300 tandem repeats of predominantly immunoglobulin-like (Ig) and fibronectin type III (FnIII) domains, with a flexible random coil-like PEVK (proline, glutamate, valine, and lysine) region. The I-

band region is primarily composed of Ig domains arranged in blocks, whereas the A-band region has alternating Ig and FnIII domains. There are single Ig-domain insertions that distinguish isoforms of titin in cardiac and skeletal muscle. Unique sequences and common splice variants are also represented: for example, the N2-A domain is expressed in all skeletal muscle isoforms, but only in the cardiac N2BA isoform. A short ~700 kDa isoform novex-3 also exists, incorporating an isoform-specific exon and is expressed in both cardiac and skeletal muscle.[220]

7.2.2 Known titin diseases

Diseases associated with mutations in titin, as of 2012 were:

Dilated cardiomyopathy 1G [221]

Familial hypertrophic cardiomyopathy 9 [222]

Early-onset myopathy with fatal cardiomyopathy [223]

Hereditary myopathy with early respiratory failure (HMERF) [224; 225]

Limb-girdle muscular dystrophy 2J (LGMD2J) [226; 227]

Tibial muscular dystrophy [227]

Later, in 2013, mutations in TTN were associated with centronuclear myopathy.[189]

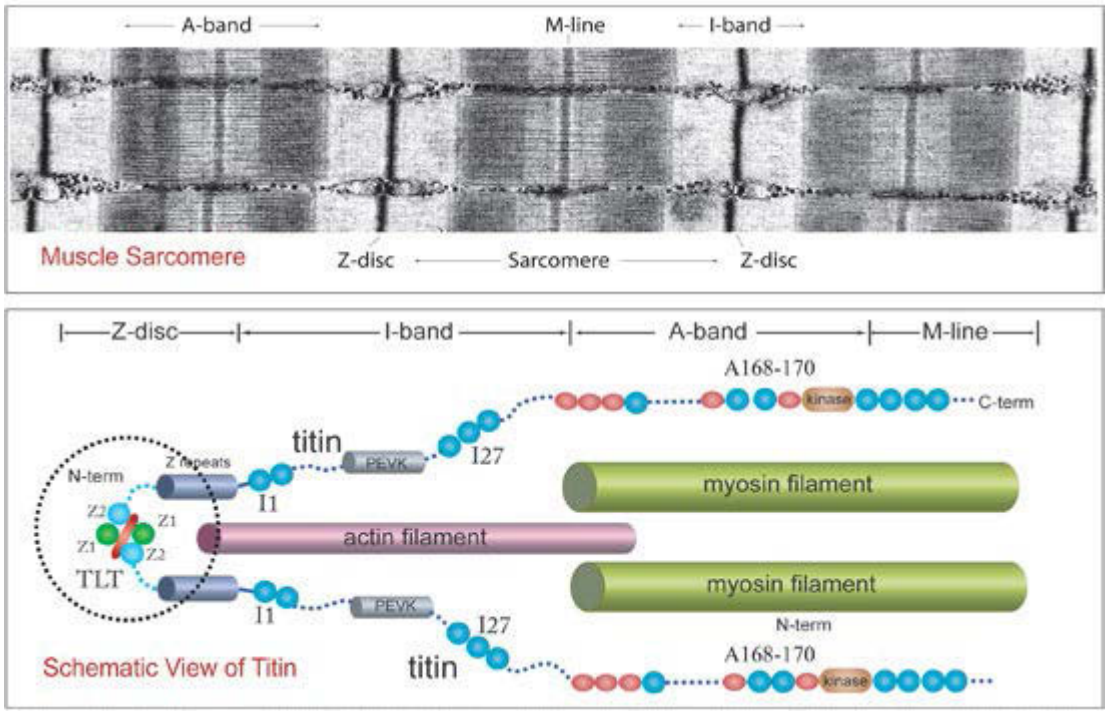


Figure 7.1: Schematic of TTN protein domains, compared with an electron micrograph of muscle sarcomeres. Regions of titin are indicated in the cartoon diagram. Images taken from Lee et al., 2007.[215] Titin spans the length of one half-sarcomere.

7.2.3 Titin skeletal muscle diseases

Early-onset myopathy with fatal cardiomyopathy was first described in two families, with a total of five affected individuals. A different homozygous frameshift deletion was identified in each family, one in exon 360, and one in exon 358. Both of these mutations were shown by immunofluorescence to truncate the TTN protein by generation of a premature stop codon, and abolish one calpain 3 (*CAPN3*) binding site. The muscle pathology for these two families also showed minicores, Type 1 fibre predominance and central nuclei.[223]

Mutations in *TTN* were first described as causative of *hereditary myopathy with early respiratory failure (HMERF)* by Lange *et al.,*2005[225] in two Swedish families and an isolated Swedish case. A single heterozygous arginine to tryptophan missense mutation was found in all affected individuals. In addition to this mutation, other missense mutations in FnIII domains in the A-band region of titin have been described as disease causing,[228; 229] as well as mutations within the titin kinase domain in the M-line. A subset of these patients also had clinical features resembling myofibrillar myopathy.[230]

The *limb-girdle muscular dystrophy 2J* (LGMD2J) phenotype was first reported in by Udd *et al.* 1992 in a large consanguineous Finnish pedigree with autosomal recessive inheritance of a severe limb-girdle muscular dystrophy. The family displayed onset of nonspecific myopathic changes on muscle biopsy and progressive fatty infiltration of involved muscles in the first to third decades, without involvement of facial muscles, or cardiomyopathy. The cause of disease in this family was found to be a homozygous 11 base-pair deletion in the last exon of *TTN*.[227]

Tibial muscular dystrophy was first described in by Udd *et al.*, as a rare adult-onset, autosomal dominant form of distal myopathy with onset in the anterior compartment of the legs.[231] Affected patients showed nonspecific dystrophic changes in affected muscles, and severe fatty replacement in the anterior tibial muscles.[231] The cause of disease in tibial muscular dystrophy was identified to be heterozygous mutations in the last exon of *TTN*. Among these heterozygous mutations was the 11 base-pair deletion found to be causative of LGMD2J when in the homozygous state.[227]

More recently, Ceyhan-Birsoy *et al.,* 2013 have described truncating and frameshift mutations within *TTN* causing *centronuclear myopathy*, with some of their patients also displaying minicores and fibre-type disproportion.[189]

7.2.4 Aims

1) To identify the causative mutations in multiminicore disease families

2) To characterise the effect of the mutations identified on the TTN protein

3) To identify a genotype-phenotype correlation between titin mutations and specific features of minicore disease.

7.3 Materials and Methods

7.3.1 Family 1 clinical features

There were three affected siblings in Family 1 (Figure 7.2). The parents were not affected and not known to be consanguineous. The phenotype was consistent across the three siblings. They presented at birth with extreme central and peripheral muscular hypotonia. There was marked delay in motor milestones, with walking only achieved after two years of age. Neck, trunk and facial weakness were moderate at two years of age, and there was moderate reduction in respiratory muscle strength. The heart was unaffected and intelligence preserved. Muscle biopsy of the quadriceps showed minicores on oxidative stains (Figure 7.3). This phenotype was not consistent with the described phenotypes of minicore disease, not displaying the severe respiratory impairment of *SEPN1* mincore disease or the ophthalmoparesis common to *RYR1* minicore disease, and a more severe phenotype than *MYH7* minicore disease, suggesting that this was a novel minicore disease phenotype.

7.3.2 Family 2 clinical features

Family 2 has a single affected individual. The unaffected parents were not known to be consanguineous. The patient presented at birth with hypotonia and weakness, reduced deep tendon reflexes and dysplastic aortic valve. Motor milestones were delayed. Sitting was achieved at 21 months and independent walking at five years of age. Muscle weakness was present in upper and lower limbs equally, with both proximal and distal involvement. Facial weakness was not present, and there was moderate reduction in respiratory muscle strength. Muscle biopsy showed multi-minicores on ATPase and oxidative stains, confirmed by electron microscopy (EM). There was marked variation in fibre size, with Type 1 fibre hypotrophy without Type 1 predominance. There was a marginal increase in central nuclei.

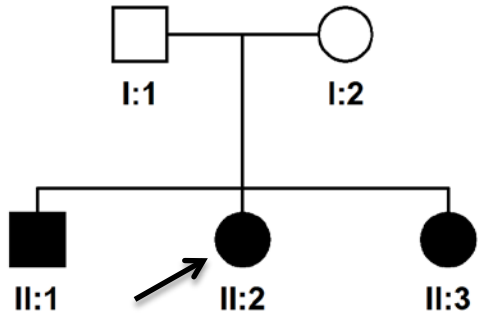


Figure 7.2: Pedigree of Family 1. Proband is indicated with an arrow.

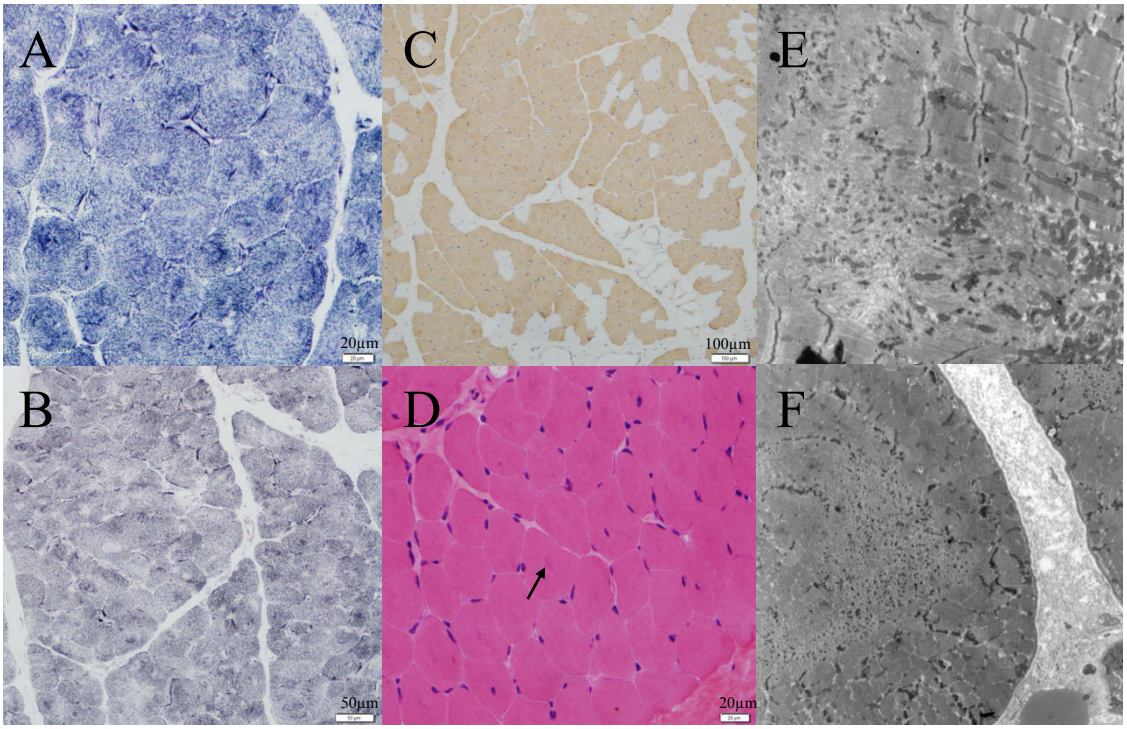


Figure 7.3: A, B: Oxidative stains of patient skeletal muscle showing cores and minicores (lighter areas). Scale bars in white. **C:** Fibre-typing (ATPase, pH 4) showing a predominance of type-1 fibres (stained fibres). Scale bar in white. **D:** H&E stain showing darker areas corresponding to cores (arrow) Scale bar in white. **E, F:** Electron micrographs of patient muscle showing ultrastructural detail of minicore regions in longitudinal (E) and transverse (F) sections.

7.3.3 Family 3 clinical features

Family 3 has a single affected individual. Similar to Family 2, the unaffected parents were not known to be consanguineous. The patient presented *in-utero* with reduced foetal movement from four months gestation, and presented at birth with bilateral flexion contractures of hips, knees, elbows and fingers. Congenital muscle weakness was not remarked upon in clinical notes. The patient presented with coarctation of the aorta that was repaired at 6 months of age, the most recent ECHO and ECG at 14 years of age were normal. Muscle weakness was present in upper and lower limbs equally, with both proximal and distal involvement. Biceps and knee reflexes were absent and brachioradialis reflexes were present but reduced. Ambulation was lost at age four years. Moderate reduction in respiratory muscle strength was present. Muscle biopsy showed large central cores and multiple smaller cores on succinate dehydrogenase (SDH) stains. These structures were not visible with ATPase staining. The presence of cores of multiple sizes was confirmed with EM. There was marked variation in fibre size, with Type 1 fibre predominance and hypotrophy. Internal nuclei were present in the majority of fibres.

7.3.4 Additional Families

In collaboration with Dr. Nigel Clark and Dr. Emily Oates (Children's Hospital at Westmead) and Dr. Volker Straub (University of Newcastle upon Tyne), an additional 8 families with minicore disease were identified and screened by Sanger, exome or targeted capture sequencing. I analysed genomic data from three of these individuals sequenced on the neurogenetic sub-exomic sequencing array described in Chapter 6. In addition to this, I performed genotype-phenotype correlation in the combined cohort using available genomic and clinical data.

7.3.5 Family 1 linkage exclusion

Linkage exclusion on sibs II:1 and II:2 in Family 1 was performed using a custom 14,514 SNP linkage map across the 22 autosomes, derived from the CytoSNP12 (Illumina) cytogenetics SNP chip. SNP data were analysed across the autosomes using the Merlin linkage suite[163] according to a recessive disease profile, following removal of unlikely genotypes and Mendelian inheritance errors. Regions of the genome with a LOD score of less than -2 were excluded, as per standard linkage analysis parameters.

7.3.6 Exome sequencing

Whole exome sequencing of individual II:2 from Family 1 was performed at the Lotterywest State Bioinformatics Facility: Genomics (LSBFG). Sequencing was performed as described in Chapter 5.3.6.

7.3.7 Neurogenetic sub-exomic sequencing array

The in-house targeted neurogenetic disease gene capture array described in Chapter 6 was used to screen 336 known neurogenetic and cardiomyopathy disease genes in single affected individuals from all three families. Sequencing of affected individuals from families 1, 2, 3 and additional external samples was performed as per the methods in Chapter 6.3.2.

7.3.8 Bioinformatic processing

Bioinformatic processing of next generation sequencing results was performed using the ANNOVAR pipeline described in chapter 3.

7.3.9 Variant confirmation

I confirmed with Sanger sequencing the variants found within Family 1 at the Western Australian Institute for Medical Research, the variants found in Families 2 and 3 were confirmed at the Neurogenetics Unit, Department of Diagnostic Genomics, PathWest Laboratory Medicine, Department of Health, Western Australia.

I designed primers for amplification of TTN exons 3, 93 and 326 for Family 1 (Primer sequences can be found in Appendix A Table 2.8). All amplicons were amplified using Qiagen HotStar™ Taq polymerase using the TD65 thermocycling protocol and standard Qiagen mix described in Appendix A.

Amplicons were sequenced using the BigDye Terminator mix (ABI) and ABI thermocycling protocol found in Appendix A. Sequencing primers were identical to the PCR amplification primers.

7.3.10 Molecular modelling of titin mutants

Molecular modelling of mutations within titin was performed using the homology modelling program Swiss-Model[232; 233] against automatically selected, pre-existing protein structures in the Swiss-Model and Protein Data Bank PDB databases.[234; 235] Visualisation of resultant protein models was performed using the PyMol program.[236]

7.4 Results

7.4.1 Linkage exclusion

Linkage exclusion analysis excluded all known core and minicore disease genes in Family 1, as well as all known myopathy disease genes save *NEB* and *TTN* on chromosome 2q.

7.4.2 Family 1

All variants identified in *TTN* are numbered according to the NM_001267550 isoform. Exome sequencing of one of the affected individuals identified two heterozygous variants in *TTN* - c.27497C>A (p.Ser9166*) in exon 95 and c.95153G>T (p.Ser31718Ile) in exon 343. Sanger sequencing confirmed the presence of both these variants in all three affected siblings, but Sanger sequencing of the parental DNA found that both variants were inherited from the mother.

Upon subsequent analysis using the neurogenetic targeted capture panel, both the variants identified previously by exome sequencing were identified. In addition, a third variant was found: an in-frame deletion of 7 base pairs (delACTAAAG) and insertion of 13 base pairs (insGGGGGGATCGATC) causing a change c.211_217delins13 (p.Thr71_Ala73delins5). This variant was confirmed by Sanger sequencing as present in all three affected siblings and shown also to be present in the unaffected father. This in-frame indel is within the highly conserved Z1 domain of TTN, known to be involved with binding to telethonin.[237; 238]

7.4.3 Family 2

DNA from the affected member of Family 2 was screened on the neurogenetic sub-exomic sequencing array only, and two heterozygous nonsense variants: c.57331C>T (p.Arg19111*) and c.105832C>T (p.Gln35278*) within *TTN* were identified and subsequently confirmed to segregate with disease by Sanger sequencing. The p.Gln35278* mutant in this family has now been described by Chauveau (2014) in a patient characterised as having dilated cardiomyopathy with minicores.[239] When I started this work in 2012, *TTN* mutations causing minicore disease had not been described.

7.4.4 Family 3

DNA from the affected member of Family 3 was screened on the neuromuscular disease gene targeted sequencing array. Two heterozygous titin variants were identified:

1) A predicted missense variant c.40558G>C (p.Val13520Leu). This variant has previously been described by Ceyhan-Birsoy *et al.,* 2013,[189] who showed that the variant caused a splice site change, leading to the skipping of exon 219.

2) A frameshift deletion c.51459_51462delTGTA (p.Asp17153Glufs*11). Sanger sequencing subsequently confirmed that these two variants segregated with disease in Family 3 one being present in the DNA of each unaffected parent.

7.4.5 Additional families

Mutations identified in 8 additional patients are summarised in Table 7.1, which also includes mutation data from families 1, 2 and 3. Within this set of 11 patients containing a total of 21 unique mutations, 19 are novel variants not previously described in literature. Two mutations had previously been described.[189; 239] A summary table of mutations can be found in Table 7.1.

7.4.6 Variant distribution

The variants found within TTN are distributed throughout most of the protein, with one variant within the Z-disk domain, 15 truncating mutations within the I- and A-band regions, two predicted non-truncating mutations within the I- and A-band regions and three truncating mutations within the M-line region. These mutations are illustrated in figure 7.4. Both the mutants in family six are C-terminal of the titin kinase domain in the M-line.

7.4.7 Modelling the possible structural effect of the disease-associated variants.

7.4.7.1 Family 1

The c.27497C>A (p.Ser9166*) in exon 95 is predicted to result in a TTN protein truncated before the PEVK region in the I-band, This mutation is predicted to result in nonsense-mediated decay.[240; 241]

Molecular modelling was performed using an existing model of the titin Z1Z2 region bound to telethonin (pdb 1YA5). The non-truncating mutation c.211_217delACTAAAGinsGGGGGGATCGATC in Family 1 abolishes a conserved alpha-helix and replaces it with 5 different amino acids. The mutation is not within the described telethonin-binding domain,[215] however, it is possible that the mutation would still destabilise the domain and disrupt the structure of this essential binding site. (A cartoon graphic of where the mutation is localised can be seen in figure 7.5.)
Homology modelling of this mutation against pdb model 1YA5 did not result in significant structural change, but this could be explained by the use of only a single template.[242]

Table 7.1: Summary table of the mutations found within the 11 studied families. All variants are numbered according to the NM_001267550 transcript, and compound heterozygous in trans unless otherwise indicated.

Family Number	Variation Type	Genomic Position	Exon	cDNA change	Predicted outcome
1	In-frame indel	chr2:179666947	Exon 3	c.211_217delACTAAAGinsGGGGGGATCGATC	p.Thr71_Ala73delins5
	Nonsense	chr2:179577152	Exon 93	c.27497C>A	p.Ser9166*
2	Nonsense	chr2:179462478	Exon 294	c.57331C>T	p.Arg19111*
	Nonsense	chr2:179395510	Exon 358	c.105832C>T^	p.Gln35278*
3	Essential splice site	chr2:179506964	Exon 219	c.40558G>C^	p.Val13520Leu, splice change, exon 219 skipped, reading frame shifted
	Frameshift deletion	chr2:179474691	Exon 272	c.51459_51462delTGTA	p.Asp17153Glufs*11
4	Nonsense	chr2:179579019	Exon 91	c.26482G>T	p.Glu8828*
	Nonsense	chr2:179463631	Exon 291	c.56806C>T	p.Arg18936*
5	Essential splice		Exon 44	c.10303+2T>C	Abolishes splice donor site, predicted frameshift
	Nonsense		Exon 162	c.35795G>T	p.Glu11932*
6	Frameshift deletion	Ch2:179391848	Exon 363	c. 107867delT	p. L35,956Rfs*16
	Essential splice	Ch2:179394685	Exon 360	c. 106375+2T>A	Predicted loss of splice donor site
7	Essential splice (Homozygous)	Chr2:179567179	Exon 108	c.30433+2T>G	Predicted loss of splice donor site
8	Essential splice	Chr2:179,499,449	Exon 230	c.42151+1G>C	Predicted loss of splice donor site
	Splice site	Chr2:179,516,177	Exon 208	c.39547+3A>G	Predicted loss of splice site, aberrant splice donor predicted at +2.
9	Frameshift deletion	Ch2:179499194	Exon 230	c.42312_42315delAAAG	p.Lys14105Asnfs*35
	Essential splice	Ch2:179444577	Exon 319	c. 67349-2A>C	Predicted to result in aberrant splicing out of exon 319 resulting in an in frame deletion of 288 bases.
10	Nonsense	Chr2:179594165	Exon 64	c.18718C>T	p.R6240*
	Essential splice	Chr2:179599054	Exon 52	c.15496+1G>T	Abolishes splice donor site, predicted frameshift
11	Frameshift	Ch2:179,466,399	Exon 287	c.55418delC	pR18473Lfs*14
	In-frame indel	chr2:179433110	Exon 326	c.77732_77749del	p.Ser25911_Tyr25917delinsTyr

^ Denotes mutations previously described in the literature

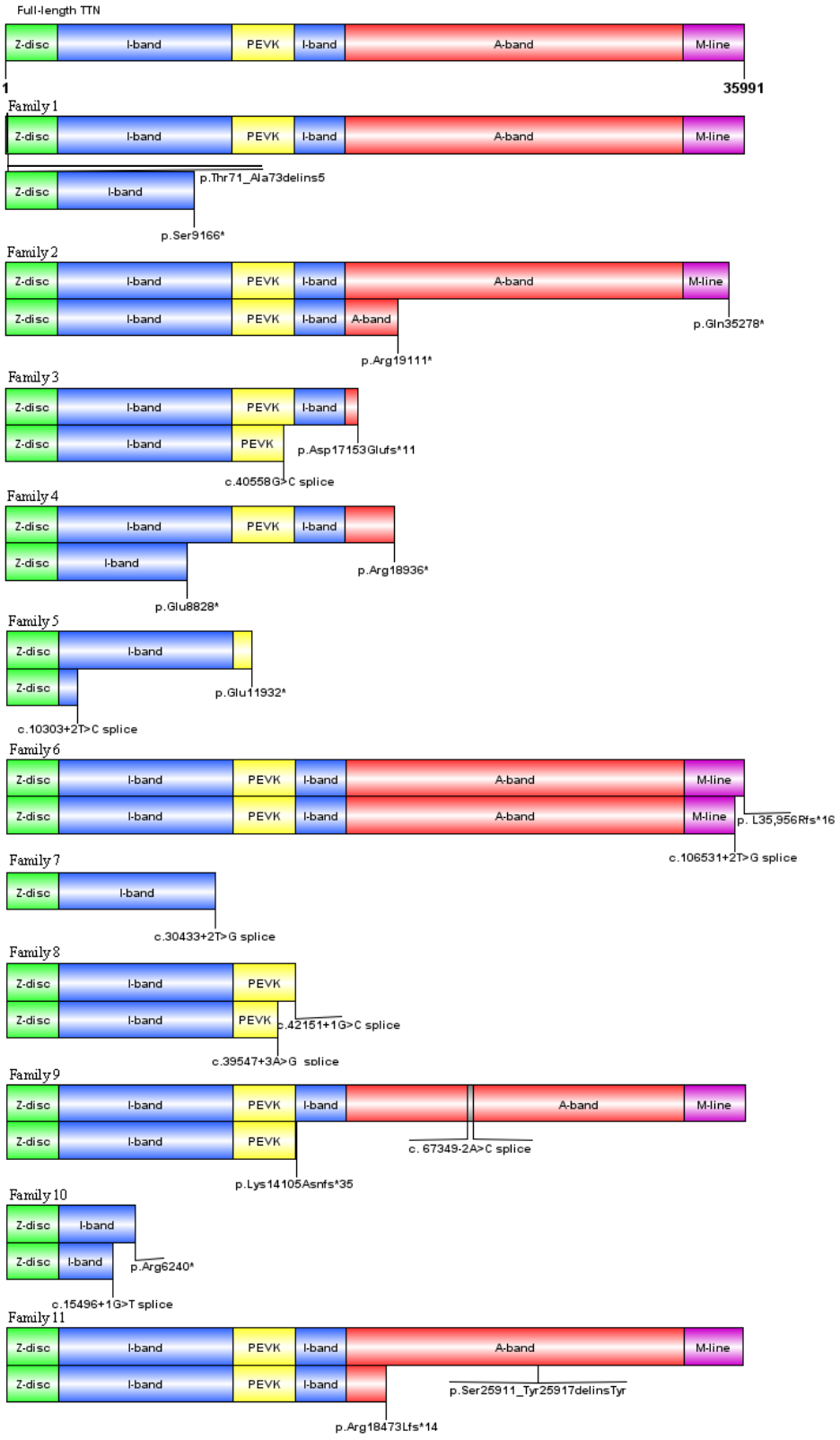


Figure 7.4: Schematic diagram of the TTN mutations found in 11 separate families, showing mutation name, position and predicted effect in the regions of the protein. Non-truncating mutations are indicated by a line leading to the location.

Figure 7.5: Cartoon of two titin Z1Z2 domains in antiparallel (teal, green) bound to a telethonin protein (purple), which is the arrangement of the protein domains in the Z-disk (pdb code 1YA5). The mutated residues in the affected alpha helix are highlighted in red, with a red arrow on both titin molecules.

7.4.7.2 Family 2

Two truncating mutations were identified in Family 2. The p.Arg19111* mutation in exon 294 is predicted to truncate the titin protein in the A-band, after the I-A junction, and the p.Gln35278* mutation in exon 358 is predicted to truncate the TTN protein before the kinase domain in the M-line. Ceyhan-Birsoy *et al.* 2013 performed immunohistochemistry on their patient cohort with truncating *TTN* mutations, and their results found that some titin protein is translated in patients where the truncating mutation is in the A-band region.[189]

7.4.7.3 Family 3

A single frameshift pathogenic mutation c.51459_51462delTGTA and a variant c.40558G>C predicted to result in aberrant splicing and a truncated protein were identified. The frameshift mutation would truncate the titin protein at the very beginning of the A-band, and the splice mutation is predicted to truncate the protein just after the PEVK region, and would possibly result in nonsense-mediated decay.

7.4.7.4 Mutation summary in the first three families

Five out of six variants identified in the three families are predicted to have truncating effects on the protein. A number of the patients with truncating TTN mutations in the centronuclear myopathy cohort examined by Ceyhan *et al.,* 2013,[189] had minicores in addition to the common pathology of internal, or central nuclei. Once segregation analysis and existing literature were considered, I considered these mutants to be disease causing in my cohort.

7.4.8 Additional families

The majority of mutations within the 8 additional patients and families analysed by Dr Emily Oates of the Children's Hospital at Westmead, are predicted to result in a truncated protein. In this patient cohort, two potential non-truncating mutations were identified. The first mutation is a splice acceptor site mutation, predicted to abolish the essential splice site at the start of exon 319 and to result in aberrant splicing and the in-frame deletion of 288 bases. Contained within this deleted region is a fibronectin III (FnIII) like domain. This mutation is within the A-band region of titin, known to bind the thick filament of the sarcomere.[218]

The second mutation is an in-frame deletion of six conserved amino acids in exon 326 (p.Ser25911_Tyr25917delinsTyr). This deletion is within a FnIII like domain.

7.4.9 Genotype-phenotype correlations

A clear correlation between age of onset and the size of the truncating mutations was not observed. In all cases where data were available, age of onset was either *in-utero*, or at birth.

Five out of 11 patients had at least one truncating mutation in the I-band region of TTN before the PEVK region. Three out of these five patients displayed congenital contractures. No clinical description of whether contractures were present at birth was available for the other two patients. A single additional patient (from Family 3) displayed congenital contractures but had compound heterozygous truncating mutations in the A-band and M-line.

Respiratory insufficiency was present in 6 out of 11 patients, with data unavailable for five.

Three patients had congenital cardiac structural abnormalities, two had cardiomyopathies, two were not seen to have cardiac abnormalities, and four patients did not have cardiac examination data.

Six out of the eleven patients developed scoliosis, with age of onset between congenital and 13 years. Data were not available for patients four and five. Of these six patients, five had two truncating mutations. In the three patients without scoliosis, one (Family one) had a non-truncating Z-line region mutation, with a predicted full-length titin protein, and a second (Family six) had two M-line truncating mutants. The third (Family nine) patient had a non-truncating mutation in the A-band. The diagnosis of scoliosis was uncertain for this patient, and was classified as negative.

A summary of these data can be seen in table 7.2.

Table 7.2: Table of genotype-phenotype correlations between the 11 patients. Mutations indicated in the mutation region columns are truncating unless otherwise stated.
FVC = Forced Vital Capacity. DCM = Dilated cardiomyopathy

Patient	Mutation 1 region	Mutation 2 region	Onset and progression	Congenital contractures	FVC	Cardiac abnormalities	Scoliosis
1	Z-disk, non-truncating	I-band before PEVK	*In-utero*, moderate	Yes	Normal	None as of last examination	No
2	A-band	M-line	*In-utero*, rapid	Yes	Abnormal, 48%, age 15	Congenital structural abnormality	Yes, age 3-14
3	PEVK	Early A-band	*In-utero*, moderate	No	Abnormal, 57%, age 5	Congenital structural abnormality	Yes, age 5
4	I-band before PEVK	Early A-band	???	???	???	???	???
5	I-band before PEVK	PEVK	At birth, slow	???	???	???	???
6	M-line	M-line	*In-utero*, slow	No	???	???	No
7	Homozygous I-band before PEVK	N/A	At birth, rapid	Yes	???	???	Yes, age 4
8	PEVK	PEVK	At birth, moderate	???	Abnormal, 41%	None as of last examination	Yes, age 13
9	I-band after PEVK	Mid A-band, non-truncating	At birth, slow	No	???	Developed DCM at age 21	No, uncertain
10	I-band before PEVK	I-band before PEVK, splice	*In-utero*, slow	Yes	Abnormal, age 12	Congenital structural abnormality	Yes, congenital
11	Early A-band	A-band, non-truncating	At birth, moderate	No	Abnormal, 40% age 9	Cardiac failure age 9	Yes, congenital

7.5 Discussion

The identification of truncating mutations in *TTN* causing recessive minicore disease expands the phenotype of *TTN* mutations beyond the currently known cardiomyopathy and myopathy phenotypes. In collaboration with a number of institutes, 21 mutations were identified, of which 19 were novel. Part of this cohort are three non-truncating, non-missense titin variants that are likely to be pathogenic.

The functional implications of homozygous and compound heterozygous titin truncating mutations in phenotypes other than severe tibial muscular dystrophy have not fully been characterised, as they are a relatively new extension of the titin disease phenotype, with truncating mutations causing cardiomyopathy described in 2012,[243] centronuclear myopathy in 2013,[189] and most recently core disease with cardiac involvement in 2014 by Chauveau *et al*.[239] However, with the addition of 20 novel mutations to the set of mutations identified from Ceyhan *et al.* and Chauveau *et al.,* further attempts can be made to investigate genotype phenotype correlations. There may be some phenotypic overlap between titin-related centronuclear myopathy and minicore disease, as core-like areas were observed in a single proband from the Ceyhan *et al.* cohort in conjunction with the primary diagnosis of centronuclear myopathy.[189] It is possible that minicore disease, centronuclear myopathy and core disease with cardiac involvement are different manifestations of the same spectrum of titin pathologies, and are caused by different combinations of truncating mutations. However, this hypothesis remains to be tested.

A clear correlation between the position of the truncating titin mutations and progression of disease could not be observed with the available information. As a comparison, patient seven has a homozygous truncating mutation that removes all of the titin protein from the PEVK region onwards, with rapid progression, and patient five has compound heterozygous truncating mutations that should result in proteins of similar size, but has slow progression. However, age of onset was consistent, either *in-utero*, or at birth.

The I-band and the PEVK region of titin are known to undergo extensive alternative splicing, which is a potential modifier for the severity of disease in patients with mutations in these regions of titin.[244] A combination of mRNA studies to identify the scope of nonsense-mediated decay, and immunohistochemistry to identify how much of each titin isoform is expressed are probably necessary to lead to a clearer correlation between mutations and disease severity.

In addition to the myopathy, it is predicted that a majority of these "titinopathy" patients will develop a cardiomyopathy in later life, due to the position of their truncating mutations. Herman *et al.,* (2012) identified a large number of heterozygous nonsense, splice and frameshift mutations in titin inferred to result in a truncated titin protein, which were associated with dominant cardiomyopathies.[243] These mutations were spread along the titin protein from the I-band region through to the A-band region, but were absent from the Z-disk and M-line regions. It is therefore likely that the patients in the cohort presented here with truncating mutations within the I and A-band regions of titin are at risk of developing a cardiomyopathy.

In the studied patient cohort, 5 of 7 patients for whom information was available (data was not available for patients 4-7) had a congenital cardiac anomaly, or developed some form of cardiac defect. Of the two patients without cardiac abnormalities, the first has double truncating mutants in the PEVK region of titin, yet no cardiac abnormalities upon their last examination at age 13. The second patient without cardiac abnormalities (last examination at age 10) has a non-truncating z-disk mutation, and a truncating mutation in the I-band, before the PEVK region. The four patients examined in the Ceyhan paper, of comparable age, also did not display any cardiac abnormalities when examined with echocardiogram and ECG.[189]

Screening and follow-up cardiac examinations should be considered for individuals with titin mutations. In the set of patients studied by Herman (2012), the mean age at diagnosis of dilated cardiomyopathy in their cohort was 37;[243] the majority of the studied cohort has not yet reached that age. Papers linking titin mutations to familial restrictive cardiomyopathy, peripartum cardiomyopathy and dilated cardiomyopathy have been published since Herman's 2012 paper.[245-247] Consequently, cardiac abnormalities should be considered as a likely clinical feature in titinopathy patients.

Congenital contractures were observed in four of eleven patients. A strong genotype-phenotype correlation cannot be drawn from this number of patients, but titin mutations should be considered when congenital contractures are observed in conjunction with a minicore phenotype.

Scoliosis should also be considered as a possible medical condition that will develop in titinopathy patients, with four patients developing it by age 13, and two presenting with congenital scoliosis.

Of note is the fact that the M-line mutation in the proband of Family 2 was described previously by Chauveau (2014) in their Patient 3. The patient in Family 2 displayed a very similar phenotype to Patient 3 in the Chaveau cohort, giving credence to Chaveau's assertion that not all truncating *TTN* mutations, particularly if in the M-line, manifest unless associated with a second mutation.[239]

The association of a further 19 novel mutations with a titinopathy phenotype expands the growing list of titin variations causing muscle disease, and will allow further genotype-phenotype correlation.

Chapter 8

Final Discussion

8.1 Final Discussion

8.1.1 Shift from classical disease gene discovery to NGS-based gene discovery

The research performed during the course of this thesis led to the identification of one novel disease gene by the classical methods of linkage analysis and positional candidate screening, and successfully combined classical and next generation disease gene discovery methods to identify a further three novel disease genes. In addition, the pipelines for continuing next generation sequencing based disease gene discovery have been established in the Harry Perkins Institute of Medical Research Neurogenetic diseases Group.

Over the course of this thesis, the fields of next generation sequencing and rare disease genetics have advanced in leaps and bounds, with more than 180 new rare disease genes discovered by exome sequencing alone between 2010 and 2013.[248]

When compared to classical disease gene discovery, exome sequencing offers a large number of benefits, offset by a known and well-characterised set of drawbacks.

The major limitations of exome sequencing are: [249-252]

1) It only is capable of screening for mutations within the exome, the protein-coding region of the genome. Deep intronic mutations and mutations in regulatory elements will go unnoticed.
2) All currently available exome sequencing techniques do not have 100% coverage of the exome – there are regions which, although targeted, are not sequenced.
3) The 'exome' as currently defined may not be every protein-coding gene, as gaps in the human sequence, and in the annotation of genes are still present.

4) Structural variations in the genome are not easily interrogated by exome sequencing – special tools and algorithms must be used, and these are inefficient.

One of the greatest benefits however, is the elimination of the need for large families from which significant linkage data can be obtained. The shift from the classical disease gene discovery method of positional candidate cloning allows the possibility of molecular diagnosis for a far greater number of families.[253] The ability that NGS has to 'cast a wider net', and bring the chance for a molecular diagnosis to a wider number of families, is in my opinion one of the greatest strengths of NGS. It is not, however, a "silver bullet", as demonstrated in Chapter 5 of this thesis; where, although three well-studied and screened families were examined using NGS, a definitive molecular diagnosis was not obtained. Decreasing costs of NGS, and the increasing prevalence of targeted gene panels serve to drive the price of sequencing per-patient down, and put this powerful tool into the reach of more patients, clinicians, diagnosticians and researchers.

The recent large-scale NGS study performed by the FORGE Canada Consortium, screened 264 disorders and identified genetic causes for 146 of them.[142] In this national project, the vast majority of results obtained were by studying trios – a mother, a father, and an affected child. Exome sequencing was performed on the trio, and the resultant variants were filtered and examined based on the inheritance, and the disease category.[142] Being the largest NGS-based disease-gene discovery project to date, it serves as a valuable model for future studies.

The combination of traditional disease gene discovery techniques, especially linkage analysis, with NGS adds more power to the analysis, reducing the number of candidate variants an analyst needs to examine. Linkage analysis is most commonly performed using single nucleotide polymorphism (SNP) arrays, but even this may change with the steadily lowering cost of exome sequencing. It may eventually be cheaper to sequence more exomes and compare the resultant data instead of performing separate SNP array based linkage analysis, especially when algorithms now exist that can perform linkage analysis using exome sequencing data, if such an analysis is desired.[254]

However, the examination of the candidate disease gene identified; the techniques used in examining the pathobiology of the mutation and the characterisation of the protein product, if the function is unknown remains the same. As such, the field of disease gene discovery may be approaching a bottleneck where the rate of disease gene *characterisation* is vastly outstripped by the pace of candidate disease gene *discovery*.

8.1.2 Bioinformatic tools and databases applied to NGS

When I started my Thesis research in 2010, very few publications describing disease genes identified using next generation sequencing (NGS) had been published[248] and NGS, even exome sequencing, was very expensive compared to 2014 prices. Furthermore, NGS facilities in Perth, Western Australia were almost non-existent, with very limited local expertise, especially in the field of bioinformatics. Because of this, all of the bioinformatic pipelines I developed in this thesis were constructed from freely available tools and programs instead of being coded in-house. I then evolved the pipelines throughout the period of my Thesis with advances in the field.

The final iteration of the bioinformatic pipeline, as described in Chapter 3 closely resembles the pipeline used by the FORGE Canada Consortium, namely being sequencer specific tools to process raw data, and the use of the ANNOVAR software and further custom scripts to annotate and filter the resultant variants.[142]

The increasing number of bioinformatic databases, and the growing number of individuals added to them allows more information to be available to researchers and clinicians. This enables more informed decisions on how important a variant is based on, for example, variant population frequency and associated phenotype data. The large scale variant databases used in this thesis were the 1000 Genomes Project (www.1000genomes.org/) and the NHLBI Exome Sequencing Project, or EVS (http://evs.gs.washington.edu/EVS/). These databases do not have phenotype data attached to the individual variants, but do have population frequency data for each variant. These databases were used to filter out 'common' variants, as defined in Chapter 3. A similar database to EVS is the Exome Aggregation Consortium (ExAC) database (http://exac.broadinstitute.org/), which is an aggregation of 61,486 unrelated individuals from various disease-specific and population studies. This database, if curated to remove affected individuals could also serve as a common variant filter.

Another such growing repository of variant data is that of the Human Variome Project.[255] First established in 2006, the project initially took a passive stance, but shifted to a more active approach, with the goal of establishing and maintaining the standards, systems and infrastructure that will enable the global collection and sharing of genetic variation information to be integrated into routine clinical practice.[256]

Future development of bioinformatic tools will hopefully allow seamless integration of more detailed variant and gene annotation databases into the variant annotation pipeline. The integration of additional informatics tools past primary variant filtering and annotation will provide a more complete NGS data analysis pathway. This will reduce the uncertainty of whether or not a variant is important, and allowing the maximum amount of information to be available to researchers and clinicians.

8.1.3 Success rates of exome sequencing-based NGS

When searching PubMed, the majority of recently published reviews on NGS, in particular, exome sequencing, have focused on the clinical utility, application, and data processing, data retention and data dissemination.[257-261] In fact, the largest published exome data set where a success rate was available was the FORGE Canada Consortium study, with a 55% success rate. Of the 146 successfully diagnosed cases, mutations were identified in 67 genes not previously associated with disease, and 42 of these have been published as of the 21st December, 2014.

The exome sequencing research performed during this thesis contributed to the identification and publication of three novel disease genes; *KLHL40*, *KHLH41* and *SPEG*,[107; 108; 262] and identified mutations in the known disease genes *TTN*, *GBE1* and *CFL2*.[191; 263] There are also two other disease genes to be published based on the NGS pipeline. Of the three neuropathy families studied in Chapter 5, candidate disease genes were identified in two families, but further gene characterisation work was halted due to lack of additional families. Calls for additional families with similar phenotypes have been published on the Australasian Neuromuscular Network (http://www.ann.org.au/) but no viable responses were received. In contrast with this, the publications for the *KLHL40*, *KHLH41* and *SPEG* disease genes were rapidly published due to availability

of samples from additional families with the same disease phenotype, in multi-institute, multinational collaborations.

With the widespread deployment of NGS technologies and the ever-increasing pace of disease gene discovery, I expect that more and more single 'interesting' families will be identified. These are families with a good candidate gene identified, but no other families with the same, or similar phenotypes immediately available for study.

As previously shown in this thesis, collaborations between existing networks of scientists and researchers can be successful, but a global initiative, instead of just local and international networks for phenotype sharing is needed to increase success rates. One such database is the PhenomeCentral repository, a secure hub for data sharing within the rare disorder community. (https://phenomecentral.org/) It allows the comparison of patient phenotype data to a repository of described phenotypes, and if a match is found, further collaborations.

8.1.4 Diagnostic applications

A major achievement of the research in this thesis is the deployment of the neurogenetic sub-exomic sequencing panel described in Chapter 6. It has become the gold standard of clinical genomics for the majority of neurogenetic diseases in Australasia, with over 900 samples screened, and a diagnostic success rate of more than a third. It has validated the Sanger fill-in approach to areas of low coverage, with five patients receiving diagnoses, and has picked up four cases where traditional Sanger screening of the gene in question did not identify the causative mutation, where the panel did.

The panel is now moving into its second iteration. The cardiomyopathy-specific disease genes have been removed from the panel, (although some cardiomyopathies can still be identified, as there is genotypic overlap with some muscle disease genes e.g. *TTN*, *MYH7*). Over 100 new neurogenetic disease genes have been added to the Version 2 panel. In addition, the concentration of capture probes targeting regions of low coverage in the version 1 panel will be increased. This 'salting' of regions of high GC content in particular is designed to increase the capture efficiency, and lower the number of regions where there are gaps in sequencing. What this technique will not do, however, is allow the mapping of short reads from repetitive regions that are targeted, like the duplicated exon clusters in *TTN* and *NEB*.

Diagnostic exomes are applied in circumstances where mutations in disease gene not on the NSES panel are suspected, many from Genetic Services Western Australia. (Personal communication, Dr. Mark Davis) When an exome is used for diagnostic purposes, there is much more data to examine – the raw variant count before filtering averages 52,000, from personal experience. Post-filtering, depending on how aggressive a strategy is used, 800 to 2,000 variants may remain, hence, incidental findings are much more likely to arise from the data. However, if a gene of interest is not on a targeted panel, and traditional Sanger sequencing would be too expensive, a diagnostic exome may be the only way to progress to a molecular diagnosis.

For diagnostic purposes, the coverage of targeted panels and the sample throughput will always be better than a diagnostic exome, and my personal opinion is that targeted panels will always remain as a first-line diagnostic screen.

8.1.5 Emerging technologies

Two emerging "second generation" NGS technologies are both single-molecule long-read sequencers, using very different technologies. The more mature technology is the Pacific Biosciences single-molecule real-time (SMRT) sequencer publically released in 2011, with the nanopore-based sequencer from Oxford Nanopore released to a small set of testing laboratories in February 2014.

The Pacific Biosciences SMRT sequencers are a fluorescent single molecule sequencing-by-synthesis technology, whereas the Oxford Nanopore sequencer uses nanopores to detect current changes when different nucleotides pass through a membrane.

Relevant to the work done in this thesis is the long read capability of these two technologies. The scale of these reads (tens of kilobases) allows researchers to look into repetitive, heterochromatic regions of the genome, such as centromeres and the Y-chromosome, as well as effectively mapping duplicated genes, and examining large-scale structural variations.[264-268] Until the technology becomes more widespread, it would be an interesting exercise to apply one of these long read technologies to the repetitive regions in the *NEB* and *TTN* genes, as the repeated exon clusters are almost identical, and their sizes (10-20kb) are amenable to sequencing as a single strand. Amplification of these regions as a single long-range PCR and subsequent sequencing would be a more effective method of sequencing these regions than Sanger fill-in. However, there is still the high error rate characteristic of these sequencers to contend with. Error rates for a single read using the Oxford Nanopore system are around 4%, as reported by Oxford Nanopore, with no third-party confirmation of these data.[269] As a system still in testing, these error rates are expected to drop to acceptable levels (<1%). The Pacific Biosciences sequencer also has a high single read error rate, but this is

offset by the well-characterised error profile, and the lack of sequencing bias in the technology. With deep enough sequencing, the final error rate after assembly approaches that of Sanger sequencing.[269; 270]

8.1.6 Final comments

It is now an extraordinary time in human genetics research, with disease genes and disease-associated genes being identified at ever-increasing pace. With this increase in sequencing availability, data analysis and retention are becoming an issue, with expertise in short supply. The information technology world is progressing at a similar, if not faster rate, with computing power growing year-on-year, and multiple huge, fast data storage methods becoming available to users outside of the enterprise market.

Wider applications for harnessing the incredible power of new genetic knowledge and technologies can now be investigated effectively and efficiently with these new technologies– next generation sequencing will allow the investigation of wide swathes of the genome around modifier genes, disease gene networks could be fully mapped, genotype-phenotype correlations could be examined and identified, all on a massive scale. However, there has been no similar revolution in the characterisation of disease genes – how and why their mutations actually cause disease, the molecular pathology of mutations. The question remains to be asked, will there be such a revolution, on the scale of next generation sequencing in that field?

One of the most important applications of next generation sequencing is the new diagnostics and prevention of genetic disease through preconception carrier screening.

The ability of both parents to know what diseases they are carriers for, and what subsequent risk that any future children may inherit a debilitating disease is in my opinion, priceless. Future advancements and economies of scale will eventually allow genetic screening to be made available to the entire population.

References

1. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.
2. (2014). Online Mendelian Inheritance in Man, OMIM®. In. (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) http://omim.org/.
3. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823-828.
4. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research 12, 996-1006.
5. Hosler, B.A., Nicholson, G.A., Sapp, P.C., Chin, W., Orrell, R.W., de Belleroche, J.S., Esteban, J., Hayward, L.J., McKenna-Yasek, D., Yeung, L., et al. (1996). Three novel mutations and two variants in the gene for Cu/Zn superoxide dismutase in familial amyotrophic lateral sclerosis. Neuromuscular disorders : NMD 6, 361-366.
6. Laing, N.G., Wilton, S.D., Akkari, P.A., Dorosz, S., Boundy, K., Kneebone, C., Blumbergs, P., White, S., Watkins, H., Love, D.R., et al. (1995). A mutation in the alpha tropomyosin gene TPM3 associated with autosomal dominant nemaline myopathy NEM1. Nature genetics 10, 249.
7. Laing, N.G., Laing, B.A., Meredith, C., Wilton, S.D., Robbins, P., Honeyman, K., Dorosz, S., Kozman, H., Mastaglia, F.L., and Kakulas, B.A. (1995). Autosomal dominant distal myopathy: linkage to chromosome 14. American journal of human genetics 56, 422-427.
8. Meredith, C., Herrmann, R., Parry, C., Liyanage, K., Dye, D.E., Durling, H.J., Duff, R.M., Beckman, K., de Visser, M., van der Graaff, M.M., et al. (2004). Mutations in the slow skeletal muscle fiber myosin heavy chain gene (MYH7) cause laing early-onset distal myopathy (MPD1). American journal of human genetics 75, 703-708.
9. Nowak, K.J., Wattanasirichaigoon, D., Goebel, H.H., Wilce, M., Pelin, K., Donner, K., Jacob, R.L., Hubner, C., Oexle, K., Anderson, J.R., et al. (1999). Mutations in the skeletal muscle alpha-actin gene in patients with actin myopathy and nemaline myopathy. Nature genetics 23, 208-212.
10. (2013). In Genome-Based Diagnostics: Demonstrating Clinical Utility in Oncology: Workshop Summary. (Washington (DC).
11. Yu, J.H., Harrell, T.M., Jamal, S.M., Tabor, H.K., and Bamshad, M.J. (2014). Attitudes of genetics professionals toward the return of incidental results from exome and whole-genome sequencing. American journal of human genetics 95, 77-84.
12. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genetics in medicine : official journal of the American College of Medical Genetics 15, 565-574.
13. Winnard, A.V., Mendell, J.R., Prior, T.W., Florence, J., and Burghes, A.H. (1995). Frameshift deletions of exons 3-7 and revertant fibers in Duchenne muscular dystrophy: mechanisms of dystrophin production. American journal of human genetics 56, 158-166.

14. Gangopadhyay, S.B., Sherratt, T.G., Heckmatt, J.Z., Dubowitz, V., Miller, G., Shokeir, M., Ray, P.N., Strong, P.N., and Worton, R.G. (1992). Dystrophin in frameshift deletion patients with Becker muscular dystrophy. American journal of human genetics 51, 562-570.
15. Lorson, C.L., Rindt, H., and Shababi, M. (2010). Spinal muscular atrophy: mechanisms and therapeutic strategies. Human molecular genetics 19, R111-118.
16. Cho, S., and Dreyfuss, G. (2010). A degron created by SMN2 exon 7 skipping is a principal contributor to spinal muscular atrophy severity. Genes & development 24, 438-442.
17. Vezain, M., Saugier-Veber, P., Goina, E., Touraine, R., Manel, V., Toutain, A., Fehrenbach, S., Frebourg, T., Pagani, F., Tosi, M., et al. (2010). A rare SMN2 variant in a previously unrecognized composite splicing regulatory element induces exon 7 inclusion and reduces the clinical severity of spinal muscular atrophy. Human mutation 31, E1110-1125.
18. Jackson, L. (1985). Prenatal genetic diagnosis by chorionic villus sampling (CVS). Seminars in perinatology 9, 209-218.
19. Kuliev, A.M., Modell, B., Jackson, L., Simpson, J.L., Brambati, B., Rhoads, G., Froster, U., Verlinsky, Y., Smidt-Jensen, S., Holzgreve, W., et al. (1992). Chorionic villus sampling (CVS): World Health Organization European Regional Office (WHO/EURO) meeting statement on the use of CVS in prenatal diagnosis. Journal of assisted reproduction and genetics 9, 299-302.
20. Biazotti, M.C., Pinto Junior, W., Albuquerque, M.C., Fujihara, L.S., Suganuma, C.H., Reigota, R.B., and Bertuzzo, C.S. (2015). Preimplantation genetic diagnosis for cystic fibrosis: a case report. Einstein 13, 110-113.
21. Barros, F.S., Araujo Junior, E., Rolo, L.C., and Nardozza, L.M. (2012). Prenatal Diagnosis of Lethal Multiple Pterygium Syndrome Using Two-and Three-Dimensional Ultrasonography. Journal of clinical imaging science 2, 65.
22. Founds, S. (2014). Innovations in prenatal genetic testing beyond the fetal karyotype. Nursing outlook.
23. Lewis, C., Hill, M., and Chitty, L.S. (2014). Non-invasive prenatal diagnosis for single gene disorders: experience of patients. Clinical genetics 85, 336-342.
24. Verlinsky, Y., Pergament, E., and Strom, C. (1990). The preimplantation genetic diagnosis of genetic diseases. Journal of in vitro fertilization and embryo transfer : IVF 7, 1-5.
25. Handyside, A.H., Kontogianni, E.H., Hardy, K., and Winston, R.M. (1990). Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. Nature 344, 768-770.
26. Helle, J.R., Braathen, G.J., Pedersen, J.C., Stokke, B., and Berg, K. (2000). [The Norwegian procedure in connection with presymptomatic testing for Huntington disease]. Tidsskrift for den Norske laegeforening : tidsskrift for praktisk medicin, ny raekke 120, 2417.
27. Scuffham, T.M., and Macmillan, J.C. (2014). Huntington Disease: Who Seeks Presymptomatic Genetic Testing, Why and What are the Outcomes? Journal of genetic counseling.
28. Eisen, A., Mezei, M.M., Stewart, H.G., Fabros, M., Gibson, G., and Andersen, P.M. (2008). SOD1 gene mutations in ALS patients from British Columbia, Canada: clinical features, neurophysiology and ethical issues in management. Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases 9, 108-119.
29. McNeil, S.M., Novelletto, A., Srinidhi, J., Barnes, G., Kornbluth, I., Altherr, M.R., Wasmuth, J.J., Gusella, J.F., MacDonald, M.E., and Myers, R.H. (1997).

Reduced penetrance of the Huntington's disease mutation. Human molecular genetics 6, 775-779.
30. Ha, A.D., and Jankovic, J. (2011). Exploring the correlates of intermediate CAG repeats in Huntington disease. Postgraduate medicine 123, 116-121.
31. Lamont, P.J., Wallefeld, W., Hilton-Jones, D., Udd, B., Argov, Z., Barboi, A.C., Bonneman, C., Boycott, K.M., Bushby, K., Connolly, A.M., et al. (2014). Novel mutations widen the phenotypic spectrum of slow skeletal/beta-cardiac myosin (MYH7) distal myopathy. Human mutation 35, 868-879.
32. Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P., Deng, H.X., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362, 59-62.
33. Sugaya, K., and Nakano, I. (2014). Prognostic role of "prion-like propagation" in SOD1-linked familial ALS: an alternative view. Frontiers in cellular neuroscience 8, 359.
34. Vehvilainen, P., Koistinaho, J., and Gundars, G. (2014). Mechanisms of mutant SOD1 induced mitochondrial toxicity in amyotrophic lateral sclerosis. Frontiers in cellular neuroscience 8, 126.
35. Hoffman, E.P., Monaco, A.P., Feener, C.C., and Kunkel, L.M. (1987). Conservation of the Duchenne muscular dystrophy gene in mice and humans. Science 238, 347-350.
36. Pelin, K., Hilpela, P., Donner, K., Sewry, C., Akkari, P.A., Wilton, S.D., Wattanasirichaigoon, D., Bang, M.L., Centner, T., Hanefeld, F., et al. (1999). Mutations in the nebulin gene associated with autosomal recessive nemaline myopathy. Proceedings of the National Academy of Sciences of the United States of America 96, 2305-2310.
37. Witt, C.C., Burkart, C., Labeit, D., McNabb, M., Wu, Y., Granzier, H., and Labeit, S. (2006). Nebulin regulates thin filament length, contractility, and Z-disk structure in vivo. The EMBO journal 25, 3843-3855.
38. Gokhin, D.S., and Fowler, V.M. (2013). A two-segment model for thin filament architecture in skeletal muscle. Nature reviews Molecular cell biology 14, 113-119.
39. Mann, C.J., Honeyman, K., Cheng, A.J., Ly, T., Lloyd, F., Fletcher, S., Morgan, J.E., Partridge, T.A., and Wilton, S.D. (2001). Antisense-induced exon skipping and synthesis of dystrophin in the mdx mouse. Proceedings of the National Academy of Sciences of the United States of America 98, 42-47.
40. Popplewell, L.J., Adkin, C., Arechavala-Gomeza, V., Aartsma-Rus, A., de Winter, C.L., Wilton, S.D., Morgan, J.E., Muntoni, F., Graham, I.R., and Dickson, G. (2010). Comparative analysis of antisense oligonucleotide sequences targeting exon 53 of the human DMD gene: Implications for future clinical trials. Neuromuscular disorders : NMD 20, 102-110.
41. Cirak, S., Arechavala-Gomeza, V., Guglieri, M., Feng, L., Torelli, S., Anthony, K., Abbs, S., Garralda, M.E., Bourke, J., Wells, D.J., et al. (2011). Exon skipping and dystrophin restoration in patients with Duchenne muscular dystrophy after systemic phosphorodiamidate morpholino oligomer treatment: an open-label, phase 2, dose-escalation study. Lancet 378, 595-605.
42. Laing, N.G., Dye, D.E., Wallgren-Pettersson, C., Richard, G., Monnier, N., Lillis, S., Winder, T.L., Lochmuller, H., Graziano, C., Mitrani-Rosenbaum, S., et al. (2009). Mutations and polymorphisms of the skeletal muscle alpha-actin gene (ACTA1). Human mutation 30, 1267-1277.
43. Nowak, K.J., Ravenscroft, G., Jackaman, C., Filipovska, A., Davies, S.M., Lim, E.M., Squire, S.E., Potter, A.C., Baker, E., Clement, S., et al. (2009). Rescue of

skeletal muscle alpha-actin-null mice by cardiac (fetal) alpha-actin. The Journal of cell biology 185, 903-915.
44. Fischer, A., Hacein-Bey, S., and Cavazzana-Calvo, M. (2002). Gene therapy of severe combined immunodeficiencies. Nature reviews Immunology 2, 615-621.
45. Denegri, M., Bongianino, R., Lodola, F., Boncompagni, S., De Giusti, V.C., Avelino-Cruz, J.E., Liu, N., Persampieri, S., Curcio, A., Esposito, F., et al. (2014). Single delivery of an adeno-associated viral construct to transfer the CASQ2 gene to knock-in mice affected by catecholaminergic polymorphic ventricular tachycardia is able to cure the disease from birth to advanced age. Circulation 129, 2673-2681.
46. Clark, J.D. (2008). The pitfalls of profoundly effective analgesic therapies. Clin J Pain 24, 825-831.
47. Morgan, T.H. (1911). The Origin of Five Mutations in Eye Color in Drosophila and Their Modes of Inheritance. Science 33, 534-537.
48. Morgan, T.H. (1915). Localization of the Hereditary Material in the Germ Cells. Proceedings of the National Academy of Sciences of the United States of America 1, 420-429.
49. Morton, N.E. (1955). Sequential tests for the detection of linkage. American journal of human genetics 7, 277-318.
50. Strachan, T., and Read, A.P. (2004). Human molecular genetics 3.(London ; New York: Garland Press).
51. Sulkowska, J.I., and Cieplak, M. (2008). Stretching to understand proteins - a survey of the protein data bank. Biophysical journal 94, 6-13.
52. Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science 265, 2037-2048.
53. Morton, N.E. (1956). The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. American journal of human genetics 8, 80-96.
54. Prescott, S.M., Lalouel, J.M., and Leppert, M. (2008). From linkage maps to quantitative trait loci: the history and science of the Utah genetic reference project. Annu Rev Genomics Hum Genet 9, 347-358.
55. Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics 32, 314-331.
56. Hui, K., Festenstein, H., de Klein, A., Grosveld, G., and Grosveld, F. (1985). HLA-DR genotyping by restriction fragment length polymorphism analyses. Immunogenetics 22, 231-239.
57. Coleman, M., Bhattacharya, S., Lindsay, S., Wright, A., Jay, M., Litt, M., Craig, I., and Davies, K. (1990). Localization of the microsatellite probe DXS426 between DXS7 and DXS255 on Xp and linkage to X-linked retinitis pigmentosa. American journal of human genetics 47, 935-940.
58. Zeggini, E., Thomson, W., Kwiatkowski, D., Richardson, A., Ollier, W., Donn, R., and British Paediatric Rheumatology Study, G. (2002). Linkage and association studies of single-nucleotide polymorphism-tagged tumor necrosis factor haplotypes in juvenile oligoarthritis. Arthritis and rheumatism 46, 3304-3311.
59. Ohashi, J., and Tokunaga, K. (2002). The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. Annals of human genetics 66, 297-306.
60. Cucca, F., Esposito, L., Goy, J.V., Merriman, M.E., Wilson, A.J., Reed, P.W., Bain, S.C., and Todd, J.A. (1998). Investigation of linkage of chromosome 8 to type 1

diabetes: multipoint analysis and exclusion mapping of human chromosome 8 in 593 affected sib-pair families from the U.K. and U.S. Diabetes 47, 1525-1527.
61. Ellison, K.A., Fill, C.P., Terwilliger, J., DeGennaro, L.J., Martin-Gallardo, A., Anvret, M., Percy, A.K., Ott, J., and Zoghbi, H. (1992). Examination of X chromosome markers in Rett syndrome: exclusion mapping with a novel variation on multilocus linkage analysis. American journal of human genetics 50, 278-287.
62. Farrer, L.A., Goodfellow, P.J., Lamarche, C.M., Franjkovic, I., Myers, S., White, B.N., Holden, J.J., Kidd, J.R., Simpson, N.E., and Kidd, K.K. (1987). An efficient strategy for gene mapping using multipoint linkage analysis: exclusion of the multiple endocrine neoplasia 2A (MEN2A) locus from chromosome 13. American journal of human genetics 40, 329-337.
63. Pagon RA, A.M., Ardinger HH, et al., editors. (1993-2014). GeneReviews® Available from: http://www.ncbi.nlm.nih.gov/books/NBK5191/. In. (University of Washington, Seattle).
64. Koenig, M., Hoffman, E.P., Bertelson, C.J., Monaco, A.P., Feener, C., and Kunkel, L.M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 50, 509-517.
65. Dubowitz, V. (1989). The Duchenne dystrophy story: from phenotype to gene and potential treatment. J Child Neurol 4, 240-250.
66. Collins, F.S. (1995). Positional cloning moves from perditional to traditional. Nature genetics 9, 347-350.
67. George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D., and Wouters, M.A. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic acids research 34, e130.
68. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.
69. Byun, M., Abhyankar, A., Lelarge, V., Plancoulaine, S., Palanduz, A., Telhan, L., Boisson, B., Picard, C., Dewell, S., Zhao, C., et al. (2010). Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. The Journal of experimental medicine 207, 2307-2312.
70. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42, 30-35.
71. Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9, 387-402.
72. Yegnasubramanian, S. (2013). Preparation of fragment libraries for next-generation sequencing on the applied biosystems SOLiD platform. Methods in enzymology 529, 185-200.
73. Bentley, D.R. (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16, 545-552.
74. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348-352.
75. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-876.

76. Mamanova, L., Coffey, A. J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat Methods 7, 111-118.
77. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. The New England journal of medicine 362, 1181-1191.
78. Hayden, E.C. (2014). Is the $1,000 genome for real? Available from: http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530. In. (Nature News, Nature Publishing Group).
79. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A., and Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6, 315-316.
80. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. (2007). Direct selection of human genomic loci by microarray hybridization. Nat Methods 4, 903-905.
81. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology 27, 182-189.
82. Campbell, C.D., and Eichler, E.E. (2013). Properties and rates of germline mutations in humans. Trends in genetics : TIG 29, 575-584.
83. Cao, M.D., Tasker, E., Willadsen, K., Imelfort, M., Vishwanathan, S., Sureshkumar, S., Balasubramanian, S., and Boden, M. (2014). Inferring short tandem repeat variation from paired-end short reads. Nucleic acids research 42, e16.
84. Suthers, G., Laing, N., Wilton, S., Dorosz, S., and Waddy, H. (1994). "Sporadic" motoneuron disease due to familial SOD1 mutation with low penetrance. Lancet 344, 1773.
85. Scacheri, P.C., Hoffman, E.P., Fratkin, J.D., Semino-Mora, C., Senchak, A., Davis, M.R., Laing, N.G., Vedanarayanan, V., and Subramony, S.H. (2000). A novel ryanodine receptor gene mutation causing both cores and rods in congenital myopathy. Neurology 55, 1689-1696.
86. Nurnberg, P., Thiele, H., Chandler, D., Hohne, W., Cunningham, M.L., Ritter, H., Leschik, G., Uhlmann, K., Mischung, C., Harrop, K., et al. (2001). Heterozygous mutations in ANKH, the human ortholog of the mouse progressive ankylosis gene, result in craniometaphyseal dysplasia. Nature genetics 28, 37-41.
87. Laing, N.G., Clarke, N.F., Dye, D.E., Liyanage, K., Walker, K.R., Kobayashi, Y., Shimakawa, S., Hagiwara, T., Ouvrier, R., Sparrow, J.C., et al. (2004). Actin mutations are one cause of congenital fibre type disproportion. Annals of neurology 56, 689-694.
88. Agrawal, P.B., Greenleaf, R.S., Tomczak, K.K., Lehtokari, V.L., Wallgren-Pettersson, C., Wallefeld, W., Laing, N.G., Darras, B.T., Maciver, S.K., Dormitzer, P.R., et al. (2007). Nemaline myopathy with minicores caused by mutation of the CFL2 gene encoding the skeletal muscle actin-binding protein, cofilin-2. Am J Hum Genet 80, 162-167.
89. Lehtokari, V.L., Ceuterick-de Groote, C., de Jonghe, P., Marttila, M., Laing, N.G., Pelin, K., and Wallgren-Pettersson, C. (2007). Cap disease caused by heterozygous deletion of the beta-tropomyosin gene TPM2. Neuromuscular disorders : NMD 17, 433-442.

90. Tsaousidou, M.K., Ouahchi, K., Warner, T.T., Yang, Y., Simpson, M.A., Laing, N.G., Wilkinson, P.A., Madrid, R.E., Patel, H., Hentati, F., et al. (2008). Sequence alterations within CYP7B1 implicate defective cholesterol homeostasis in motor-neuron degeneration. American journal of human genetics 82, 510-515.
91. Gommans, I.M., van Engelen, B.G., ter Laak, H.J., Brunner, H.G., Kremer, H., Lammens, M., and Vogels, O.J. (2002). A new phenotype of autosomal dominant nemaline myopathy. Neuromuscular disorders : NMD 12, 13-18.
92. Gommans, I.M., Davis, M., Saar, K., Lammens, M., Mastaglia, F., Lamont, P., van Duijnhoven, G., ter Laak, H.J., Reis, A., Vogels, O.J., et al. (2003). A locus on chromosome 15q for a dominantly inherited nemaline myopathy with core-like lesions. Brain : a journal of neurology 126, 1545-1551.
93. Pauw-Gommans, I.M., Gerrits, K.H., de Haan, A., and van Engelen, B.G. (2006). Muscle slowness in a family with nemaline myopathy. Neuromuscular disorders : NMD 16, 477-480.
94. Olive, M., Goldfarb, L.G., Lee, H.S., Odgerel, Z., Blokhin, A., Gonzalez-Mera, L., Moreno, D., Laing, N.G., and Sambuughin, N. (2010). Nemaline myopathy type 6: clinical and myopathological features. Muscle & nerve 42, 901-907.
95. Sambuughin, N., Yau, K.S., Olive, M., Duff, R.M., Bayarsaikhan, M., Lu, S., Gonzalez-Mera, L., Sivadorai, P., Nowak, K.J., Ravenscroft, G., et al. (2010). Dominant mutations in KBTBD13, a member of the BTB/Kelch family, cause nemaline myopathy with cores. American journal of human genetics 87, 842-847.
96. Bork, P., and Doolittle, R.F. (1994). Drosophila kelch motif is derived from a common enzyme fold. Journal of molecular biology 236, 1277-1282.
97. Xue, F., and Cooley, L. (1993). kelch encodes a component of intercellular bridges in Drosophila egg chambers. Cell 72, 681-693.
98. Cirak, S., von Deimling, F., Sachdev, S., Errington, W.J., Herrmann, R., Bonnemann, C., Brockmann, K., Hinderlich, S., Lindner, T.H., Steinbrecher, A., et al. (2010). Kelch-like homologue 9 mutation is associated with an early onset autosomal dominant distal myopathy. Brain : a journal of neurology 133, 2123-2135.
99. Brownstein, M.J., Carpten, J.D., and Smith, J.R. (1996). Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. BioTechniques 20, 1004-1006, 1008-1010.
100. De Marco, V., Stier, G., Blandin, S., and de Marco, A. (2004). The solubility and stability of recombinant proteins are increased by their fusion to NusA. Biochemical and biophysical research communications 322, 766-771.
101. Quan, S., Koldewey, P., Tapley, T., Kirsch, N., Ruane, K.M., Pfizenmaier, J., Shi, R., Hofmann, S., Foit, L., Ren, G., et al. (2011). Genetic selection designed to stabilize proteins uncovers a chaperone called Spy. Nature structural & molecular biology 18, 262-269.
102. Romero, N.B., Sandaradura, S.A., and Clarke, N.F. (2013). Recent advances in nemaline myopathy. Current opinion in neurology 26, 519-526.
103. Sambuughin, N., Swietnicki, W., Techtmann, S., Matrosova, V., Wallace, T., Goldfarb, L., and Maynard, E. (2012). KBTBD13 interacts with Cullin 3 to form a functional ubiquitin ligase. Biochemical and biophysical research communications 421, 743-749.
104. Adams, J., Kelso, R., and Cooley, L. (2000). The kelch repeat superfamily of proteins: propellers of cell function. Trends in cell biology 10, 17-24.

105. Pessah, I.N., Waterhouse, A.L., and Casida, J.E. (1985). The calcium-ryanodine receptor complex of skeletal and cardiac muscle. Biochemical and biophysical research communications 128, 449-456.
106. Jurynec, M.J., Xia, R., Mackrill, J.J., Gunther, D., Crawford, T., Flanigan, K.M., Abramson, J.J., Howard, M.T., and Grunwald, D.J. (2008). Selenoprotein N is required for ryanodine receptor calcium release channel activity in human and zebrafish muscle. Proceedings of the National Academy of Sciences of the United States of America 105, 12485-12490.
107. Ravenscroft, G., Miyatake, S., Lehtokari, V.L., Todd, E.J., Vornanen, P., Yau, K.S., Hayashi, Y.K., Miyake, N., Tsurusaki, Y., Doi, H., et al. (2013). Mutations in KLHL40 are a frequent cause of severe autosomal-recessive nemaline myopathy. American journal of human genetics 93, 6-18.
108. Gupta, V.A., Ravenscroft, G., Shaheen, R., Todd, E.J., Swanson, L.C., Shiina, M., Ogata, K., Hsu, C., Clarke, N.F., Darras, B.T., et al. (2013). Identification of KLHL41 Mutations Implicates BTB-Kelch-Mediated Ubiquitination as an Alternate Pathway to Myofibrillar Disruption in Nemaline Myopathy. American journal of human genetics 93, 1108-1117.
109. Garg, A., O'Rourke, J., Long, C., Doering, J., Ravenscroft, G., Bezprozvannaya, S., Nelson, B.R., Beetz, N., Li, L., Chen, S., et al. (2014). KLHL40 deficiency destabilizes thin filament proteins and promotes nemaline myopathy. The Journal of clinical investigation 124, 3529-3539.
110. Zhang, D.D., Lo, S.C., Sun, Z., Habib, G.M., Lieberman, M.W., and Hannink, M. (2005). Ubiquitination of Keap1, a BTB-Kelch substrate adaptor protein for Cul3, targets Keap1 for degradation by a proteasome-independent pathway. The Journal of biological chemistry 280, 30091-30099.
111. Subramaniam, S. (2004). Bioinformatics and computational systems biology: at the cross roads of biology, engineering and computation. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference 7, 5458.
112. Gullapalli, R.R., Lyons-Weiler, M., Petrosko, P., Dhir, R., Becich, M.J., and LaFramboise, W.A. (2012). Clinical integration of next-generation sequencing technology. Clinics in laboratory medicine 32, 585-599.
113. Chen, W., Kalscheuer, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M.H., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G., et al. (2008). Mapping translocation breakpoints by next-generation sequencing. Genome research 18, 1143-1149.
114. Millat, G., Chanavat, V., and Rousson, R. (2014). Evaluation of a New High-Throughput Next-Generation Sequencing Method Based on a Custom AmpliSeq Library and Ion Torrent PGM Sequencing for the Rapid Detection of Genetic Variations in Long QT Syndrome. Molecular diagnosis & therapy.
115. Singh, R.R., Patel, K.P., Routbort, M.J., Reddy, N.G., Barkoh, B.A., Handal, B., Kanagal-Shamanna, R., Greaves, W.O., Medeiros, L.J., Aldape, K.D., et al. (2013). Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. The Journal of molecular diagnostics : JMD 15, 607-622.
116. Lin, B., Wang, J., and Cheng, Y. (2008). Recent Patents and Advances in the Next-Generation Sequencing Technologies. Recent patents on biomedical engineering 2008, 60-67.
117. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.

118. Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC bioinformatics 11, 485.
119. Boland, J.F., Chung, C.C., Roberson, D., Mitchell, J., Zhang, X., Im, K.M., He, J., Chanock, S.J., Yeager, M., and Dean, M. (2013). The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. Human genetics 132, 1153-1163.
120. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Proc, G.P.D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
121. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
122. Leung, R.K., Dong, Z.Q., Sa, F., Chong, C.M., Lei, S.W., Tsui, S.K., and Lee, S.M. (2014). Quick, sensitive and specific detection and evaluation of quantification of minor variants by high-throughput sequencing. Molecular bioSystems 10, 206-214.
123. Susan Tang, F.C.L.H., Thomas C. Wessel, Jon Sorenson, Heather Peckham, Francisco M. De La Vega. (2008). DiBayes: A SNP Detection Algorithm for Next-Generation Dibase Sequencing. In. (Applied Biosystems, Foster City, CA, and Beverly, MA, USA.
124. Dolled-Filhart, M.P., Lee, M., Jr., Ou-Yang, C.W., Haraksingh, R.R., and Lin, J.C. (2013). Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. TheScientificWorldJournal 2013, 730210.
125. Dogan, H., Can, H., and Otu, H.H. (2014). Whole genome sequence of a Turkish individual. PloS one 9, e85233.
126. Robinson, P.N., Kohler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome research 24, 340-348.
127. Tranchevent, L.C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic acids research 36, W377-384.
128. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164.
129. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.
130. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
131. Ge, D., Ruzzo, E.K., Shianna, K.V., He, M., Pelak, K., Heinzen, E.L., Need, A.C., Cirulli, E.T., Maia, J.M., Dickson, S.P., et al. (2011). SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics 27, 1998-2000.
132. Sherry, S.T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome research 9, 677-679.

133. Smigielski, E.M., Sirotkin, K., Ward, M., and Sherry, S.T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic acids research 28, 352-355.
134. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.
135. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760-1774.
136. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
137. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/). In. (
138. Shi, Y., and Majewski, J. (2013). FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. Bioinformatics 29, 1461-1462.
139. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 14, 178-192.
140. Buske, O.J., Manickaraj, A., Mital, S., Ray, P.N., and Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. Bioinformatics 29, 1843-1850.
141. Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., Jin, X., Jiang, H., Zhang, P., Shen, L., Guo, J.F., et al. (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain : a journal of neurology 133, 3510-3518.
142. Beaulieu, C.L., Majewski, J., Schwartzentruber, J., Samuels, M.E., Fernandez, B.A., Bernier, F.P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. American journal of human genetics 94, 809-817.
143. Dieterich, K., Quijano-Roy, S., Monnier, N., Zhou, J., Faure, J., Smirnow, D.A., Carlier, R., Laroche, C., Marcorelles, P., Mercier, S., et al. (2013). The neuronal endopeptidase ECEL1 is associated with a distinct form of recessive distal arthrogryposis. Human molecular genetics 22, 1483-1492.
144. McMillin, M.J., Below, J.E., Shively, K.M., Beck, A.E., Gildersleeve, H.I., Pinner, J., Gogola, G.R., Hecht, J.T., Grange, D.K., Harris, D.J., et al. (2013). Mutations in ECEL1 cause distal arthrogryposis type 5D. American journal of human genetics 92, 150-156.
145. Shaheen, R., Al-Owain, M., Khan, A.O., Zaki, M.S., Hossni, H.A., Al-Tassan, R., Eyaid, W., and Alkuraya, F.S. (2014). Identification of three novel ECEL1 mutations in three families with distal arthrogryposis type 5D. Clinical genetics 85, 568-572.
146. Ockeloen, C.W., Gilhuis, H.J., Pfundt, R., Kamsteeg, E.J., Agrawal, P.B., Beggs, A.H., Dara Hama-Amin, A., Diekstra, A., Knoers, N.V., Lammens, M., et al. (2012). Congenital myopathy caused by a novel missense mutation in the CFL2 gene. Neuromuscular disorders : NMD 22, 632-639.
147. Klein, C.J., Duan, X., and Shy, M.E. (2013). Inherited neuropathies: clinical overview and update. Muscle & nerve 48, 604-622.

148. Fink, J.K. (2013). Hereditary spastic paraplegia: clinico-pathologic features and emerging molecular mechanisms. Acta neuropathologica 126, 307-328.
149. Lo Giudice, T., Lombardi, F., Santorelli, F.M., Kawarai, T., and Orlacchio, A. (2014). Hereditary spastic paraplegia: Clinical-genetic characteristics and evolving molecular mechanisms. Experimental neurology 261, 518-539.
150. Schule, R., and Schols, L. (2011). Genetics of hereditary spastic paraplegias. Seminars in neurology 31, 484-493.
151. Finsterer, J., Loscher, W., Quasthoff, S., Wanschitz, J., Auer-Grumbach, M., and Stevanin, G. (2012). Hereditary spastic paraplegias with autosomal dominant, recessive, X-linked, or maternal trait of inheritance. Journal of the neurological sciences 318, 1-18.
152. Ruano, L., Melo, C., Silva, M.C., and Coutinho, P. (2014). The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. Neuroepidemiology 42, 174-183.
153. Ciccarelli, F.D., Proukakis, C., Patel, H., Cross, H., Azam, S., Patton, M.A., Bork, P., and Crosby, A.H. (2003). The identification of a conserved domain in both spartin and spastin, mutated in hereditary spastic paraplegia. Genomics 81, 437-441.
154. Sanderson, C.M., Connell, J.W., Edwards, T.L., Bright, N.A., Duley, S., Thompson, A., Luzio, J.P., and Reid, E. (2006). Spastin and atlastin, two proteins mutated in autosomal-dominant hereditary spastic paraplegia, are binding partners. Human molecular genetics 15, 307-318.
155. Novarino, G., Fenstermaker, A.G., Zaki, M.S., Hofree, M., Silhavy, J.L., Heiberg, A.D., Abdellateef, M., Rosti, B., Scott, E., Mansour, L., et al. (2014). Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. Science 343, 506-511.
156. Durr, A. (2010). Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. Lancet neurology 9, 885-894.
157. Harding, A.E. (1982). The clinical features and classification of the late onset autosomal dominant cerebellar ataxias. A study of 11 families, including descendants of the 'the Drew family of Walworth'. Brain : a journal of neurology 105, 1-28.
158. Schols, L., Bauer, P., Schmidt, T., Schulte, T., and Riess, O. (2004). Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. Lancet neurology 3, 291-304.
159. Metz, G., Coppard, N., Cooper, J.M., Delatycki, M.B., Durr, A., Di Prospero, N.A., Giunti, P., Lynch, D.R., Schulz, J.B., Rummey, C., et al. (2013). Rating disease progression of Friedreich's ataxia by the International Cooperative Ataxia Rating Scale: analysis of a 603-patient database. Brain : a journal of neurology 136, 259-268.
160. Fogel, B.L., and Perlman, S. (2007). Clinical features and molecular genetics of autosomal recessive cerebellar ataxias. Lancet neurology 6, 245-257.
161. Anheim, M., Tranchant, C., and Koenig, M. (2012). The autosomal recessive cerebellar ataxias. The New England journal of medicine 366, 636-646.
162. Palau, F., and Espinos, C. (2006). Autosomal recessive cerebellar ataxias. Orphanet journal of rare diseases 1, 47.
163. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30, 97-101.
164. Ruschendorf, F., and Nurnberg, P. (2005). ALOHOMORA: a tool for linkage analysis using 10K SNP array data. Bioinformatics 21, 2123-2125.

165. Henneke, M., Wehner, L.E., Hennies, H.C., Preuss, N., and Gartner, J. (2004). Mutation analysis of the M6b gene in patients with Pelizaeus-Merzbacher-like syndrome. American journal of medical genetics Part A 128A, 156-158.
166. Gardinier, M.V., and Macklin, W.B. (1988). Myelin proteolipid protein gene expression in jimpy and jimpy(msd) mice. Journal of neurochemistry 51, 360-369.
167. Werner, H.B., Kramer-Albers, E.M., Strenzke, N., Saher, G., Tenzer, S., Ohno-Iwashita, Y., De Monasterio-Schrader, P., Mobius, W., Moser, T., Griffiths, I.R., et al. (2013). A critical role for the cholesterol-associated proteolipids PLP and M6B in myelination of the central nervous system. Glia 61, 567-586.
168. Daulat, A.M., Maurice, P., Froment, C., Guillaume, J.L., Broussard, C., Monsarrat, B., Delagrange, P., and Jockers, R. (2007). Purification and identification of G protein-coupled receptor protein complexes under native conditions. Molecular & cellular proteomics : MCP 6, 835-844.
169. Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell 138, 389-403.
170. Matsumoto, M., Hatakeyama, S., Oyamada, K., Oda, Y., Nishimura, T., and Nakayama, K.I. (2005). Large-scale analysis of the human ubiquitin-related proteome. Proteomics 5, 4145-4151.
171. Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al. (2011). A protein complex network of Drosophila melanogaster. Cell 147, 690-703.
172. MacArthur, M.W., and Thornton, J.M. (1991). Influence of proline residues on protein conformation. Journal of molecular biology 218, 397-412.
173. Dagvadorj, A., Goudeau, B., Hilton-Jones, D., Blancato, J.K., Shatunov, A., Simon-Casteras, M., Squier, W., Nagle, J.W., Goldfarb, L.G., and Vicart, P. (2003). Respiratory insufficiency in desminopathy patients caused by introduction of proline residues in desmin c-terminal alpha-helical segment. Muscle & nerve 27, 669-675.
174. Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. Jama 312, 1880-1887.
175. Lynch, D.R., and Farmer, J. (2002). Practical approaches to neurogenetic disease. Journal of neuro-ophthalmology : the official journal of the North American Neuro-Ophthalmology Society 22, 297-304.
176. Mazzocco, M.l.M.M., and Ross, J.L. (2007). Neurogenetic developmental disorders : variation of manifestation in childhood.(Cambridge, Mass.: MIT Press).
177. Kaplan, J.C. (2011). The 2012 version of the gene table of monogenic neuromuscular disorders. Neuromuscular disorders : NMD 21, 833-861.
178. Laing, N.G. (2012). Genetics of neuromuscular disorders. Critical reviews in clinical laboratory sciences 49, 33-48.
179. Kang, P.B. (2013). Ethical issues in neurogenetic disorders. Handbook of clinical neurology 118, 265-276.
180. Carboni, N., Mateddu, A., Marrosu, G., Cocco, E., and Marrosu, M.G. (2013). Genetic and clinical characteristics of skeletal and cardiac muscle in patients with lamin A/C gene mutations. Muscle & nerve 48, 161-170.
181. Wilcox, R., Braenne, I., Bruggemann, N., Winkler, S., Wiegers, K., Bertram, L., Anderson, T., and Lohmann, K. (2014). Genome sequencing identifies a novel mutation in ATP1A3 in a family with dystonia in females only. Journal of neurology.

182. Delanty, N., and Goldstein, D.B. (2013). Diagnostic exome sequencing: a new paradigm in neurology. Neuron 80, 841-843.
183. Nigro, V., and Piluso, G. (2012). Next generation sequencing (NGS) strategies for the genetic testing of myopathies. Acta myologica : myopathies and cardiomyopathies : official journal of the Mediterranean Society of Myology / edited by the Gaetano Conte Academy for the study of striated muscle diseases 31, 196-200.
184. Leidenroth, A., Sorte, H.S., Gilfillan, G., Ehrlich, M., Lyle, R., and Hewitt, J.E. (2012). Diagnosis by sequencing: correction of misdiagnosis from FSHD2 to LGMD2A by whole-exome analysis. European journal of human genetics : EJHG 20, 999-1003.
185. Valencia, C.A., Ankala, A., Rhodenizer, D., Bhide, S., Littlejohn, M.R., Keong, L.M., Rutkowski, A., Sparks, S., Bonnemann, C., and Hegde, M. (2013). Comprehensive mutation analysis for congenital muscular dystrophy: a clinical PCR-based enrichment and next-generation sequencing panel. PloS one 8, e53083.
186. Vasli, N., Bohm, J., Le Gras, S., Muller, J., Pizot, C., Jost, B., Echaniz-Laguna, A., Laugel, V., Tranchant, C., Bernard, R., et al. (2012). Next generation sequencing for molecular diagnosis of neuromuscular diseases. Acta neuropathologica 124, 273-283.
187. Oldfors, A., and Lamont, P.J. (2008). Thick filament diseases. Advances in experimental medicine and biology 642, 78-91.
188. LeWinter, M.M., and Granzier, H.L. (2013). Titin is a major human disease gene. Circulation 127, 938-944.
189. Ceyhan-Birsoy, O., Agrawal, P.B., Hidalgo, C., Schmitz-Abe, K., DeChene, E.T., Swanson, L.C., Soemedi, R., Vasli, N., Iannaccone, S.T., Shieh, P.B., et al. (2013). Recessive truncating titin gene, TTN, mutations presenting as centronuclear myopathy. Neurology 81, 1205-1214.
190. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.
191. Ravenscroft, G., Thompson, E.M., Todd, E.J., Yau, K.S., Kresoje, N., Sivadorai, P., Friend, K., Riley, K., Manton, N.D., Blumbergs, P., et al. (2013). Whole exome sequencing in foetal akinesia expands the genotype-phenotype spectrum of GBE1 glycogen storage disease mutations. Neuromuscular disorders : NMD 23, 165-169.
192. Chilamakuri, C.S., Lorenz, S., Madoui, M.A., Vodak, D., Sun, J., Hovig, E., Myklebost, O., and Meza-Zepeda, L.A. (2014). Performance comparison of four exome capture systems for deep sequencing. BMC genomics 15, 449.
193. Ye, Y., Yu, P., Yong, J., Zhang, T., Wei, X., Qi, M., and Jin, F. (2014). Preimplantational Genetic Diagnosis and Mutation Detection in a Family with Duplication Mutation of DMD Gene. Gynecologic and obstetric investigation.
194. Graziano, C., Bertini, E., Minetti, C., and Porfirio, B. (2004). Alpha-actin gene mutations and polymorphisms in Italian patients with nemaline myopathy. International journal of molecular medicine 13, 805-809.
195. Ilkovski, B., Cooper, S.T., Nowak, K., Ryan, M.M., Yang, N., Schnell, C., Durling, H.J., Roddick, L.G., Wilkinson, I., Kornberg, A.J., et al. (2001). Nemaline myopathy caused by mutations in the muscle alpha-skeletal-actin gene. American journal of human genetics 68, 1333-1343.
196. Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L.H., Kamsteeg, E.J., Mensenkamp, A.R., Rodenburg, R.J., Yntema, H.G., Spruijt, L., Vermeer, S., et al. (2013). A post-hoc comparison of the utility of sanger sequencing and exome

sequencing for the diagnosis of heterogeneous diseases. Human mutation 34, 1721-1726.
197. Lepri, F.R., Scavelli, R., Digilio, M.C., Gnazzo, M., Grotta, S., Dentici, M.L., Pisaneschi, E., Sirleto, P., Capolino, R., Baban, A., et al. (2014). Diagnosis of Noonan syndrome and related disorders using target next generation sequencing. BMC medical genetics 15, 14.
198. Besnard, T., Garcia-Garcia, G., Baux, D., Vache, C., Faugere, V., Larrieu, L., Leonard, S., Millan, J.M., Malcolm, S., Claustres, M., et al. (2014). Experience of targeted Usher exome sequencing as a clinical test. Molecular genetics & genomic medicine 2, 30-43.
199. Nemeth, A.H., Kwasniewska, A.C., Lise, S., Parolin Schnekenberg, R., Becker, E.B., Bera, K.D., Shanks, M.E., Gregory, L., Buck, D., Zameel Cader, M., et al. (2013). Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. Brain : a journal of neurology 136, 3106-3118.
200. Brett, M., McPherson, J., Zang, Z.J., Lai, A., Tan, E.S., Ng, I., Ong, L.C., Cham, B., Tan, P., Rozen, S., et al. (2014). Massively parallel sequencing of patients with intellectual disability, congenital anomalies and/or autism spectrum disorders with a targeted gene panel. PloS one 9, e93409.
201. Glockle, N., Kohl, S., Mohr, J., Scheurenbrand, T., Sprecher, A., Weisschuh, N., Bernd, A., Rudolph, G., Schubach, M., Poloschek, C., et al. (2013). Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies. European journal of human genetics : EJHG.
202. Sulonen, A.M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., et al. (2011). Comparison of solution-based exome capture methods for next generation sequencing. Genome biology 12, R94.
203. Coonrod, E.M., Durtschi, J.D., Webb, C.V., Voelkerding, K.V., and Kumanovics, A. (2014). Next-generation sequencing of custom amplicons to improve coverage of HaloPlex multigene panels. BioTechniques 57, 204-207.
204. Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature reviews Genetics 12, 683-691.
205. Wei, X., Dai, Y., Yu, P., Qu, N., Lan, Z., Hong, X., Sun, Y., Yang, G., Xie, S., Shi, Q., et al. (2013). Targeted next-generation sequencing as a comprehensive test for patients with and female carriers of DMD/BMD: a multi-population diagnostic study. European journal of human genetics : EJHG.
206. Eisenberger, T., Neuhaus, C., Khan, A.O., Decker, C., Preising, M.N., Friedburg, C., Bieg, A., Gliem, M., Charbel Issa, P., Holz, F.G., et al. (2013). Increasing the yield in targeted next-generation sequencing by implicating CNV analysis, non-coding exons and the overall variant load: the example of retinal dystrophies. PloS one 8, e78496.
207. Cullup, T., Lamont, P.J., Cirak, S., Damian, M.S., Wallefeld, W., Gooding, R., Tan, S.V., Sheehan, J., Muntoni, F., Abbs, S., et al. (2012). Mutations in MYH7 cause Multi-minicore Disease (MmD) with variable cardiac involvement. Neuromuscular disorders : NMD 22, 1096-1104.
208. Zhou, H., Jungbluth, H., Sewry, C.A., Feng, L., Bertini, E., Bushby, K., Straub, V., Roper, H., Rose, M.R., Brockington, M., et al. (2007). Molecular mechanisms and phenotypic variation in RYR1-related congenital myopathies. Brain : a journal of neurology 130, 2024-2036.
209. Ferreiro, A., Quijano-Roy, S., Pichereau, C., Moghadaszadeh, B., Goemans, N., Bonnemann, C., Jungbluth, H., Straub, V., Villanova, M., Leroy, J.P., et al.

(2002). Mutations of the selenoprotein N gene, which is implicated in rigid spine muscular dystrophy, cause the classical phenotype of multiminicore disease: reassessing the nosology of early-onset myopathies. American journal of human genetics 71, 739-749.
210. Maiti, B., Arbogast, S., Allamand, V., Moyle, M.W., Anderson, C.B., Richard, P., Guicheney, P., Ferreiro, A., Flanigan, K.M., and Howard, M.T. (2009). A mutation in the SEPN1 selenocysteine redefinition element (SRE) reduces selenocysteine incorporation and leads to SEPN1-related myopathy. Human mutation 30, 411-416.
211. Zhang, Y., Chen, H.S., Khanna, V.K., De Leon, S., Phillips, M.S., Schappert, K., Britt, B.A., Browell, A.K., and MacLennan, D.H. (1993). A mutation in the human ryanodine receptor gene associated with central core disease. Nature genetics 5, 46-50.
212. Scoto, M., Cullup, T., Cirak, S., Yau, S., Manzur, A.Y., Feng, L., Jacques, T.S., Anderson, G., Abbs, S., Sewry, C., et al. (2013). Nebulin (NEB) mutations in a childhood onset distal myopathy with rods and cores uncovered by next generation sequencing. European journal of human genetics : EJHG 21, 1249-1252.
213. Tskhovrebova, L., Trinick, J., Sleep, J.A., and Simmons, R.M. (1997). Elasticity and unfolding of single molecules of the giant muscle protein titin. Nature 387, 308-312.
214. Tskhovrebova, L., and Trinick, J. (1997). Direct visualization of extensibility in isolated titin molecules. Journal of molecular biology 265, 100-106.
215. Lee, E.H., Hsin, J., Mayans, O., and Schulten, K. (2007). Secondary and tertiary structure elasticity of titin Z1Z2 and a titin chain model. Biophysical journal 93, 1719-1735.
216. Linke, W.A., Rudy, D.E., Centner, T., Gautel, M., Witt, C., Labeit, S., and Gregorio, C.C. (1999). I-band titin in cardiac muscle is a three-element molecular spring and is critical for maintaining thin filament structure. The Journal of cell biology 146, 631-644.
217. Soteriou, A., Gamage, M., and Trinick, J. (1993). A survey of interactions made by the giant protein titin. Journal of cell science 104 (Pt 1), 119-123.
218. Kruger, M., and Linke, W.A. (2011). The giant protein titin: a regulatory node that integrates myocyte signaling pathways. The Journal of biological chemistry 286, 9905-9912.
219. McElhinny, A.S., Kakinuma, K., Sorimachi, H., Labeit, S., and Gregorio, C.C. (2002). Muscle-specific RING finger-1 interacts with titin to regulate sarcomeric M-line and thick filament structure and may have nuclear functions via its interaction with glucocorticoid modulatory element binding protein-1. The Journal of cell biology 157, 125-136.
220. Bang, M.L., Centner, T., Fornoff, F., Geach, A.J., Gotthardt, M., McNabb, M., Witt, C.C., Labeit, D., Gregorio, C.C., Granzier, H., et al. (2001). The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. Circulation research 89, 1065-1072.
221. Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitas, K., Sasse-Klaassen, S., Seidman, J.G., Seidman, C., Granzier, H., Labeit, S., et al. (2002). Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. Nature genetics 30, 201-204.
222. Satoh, M., Takahashi, M., Sakamoto, T., Hiroe, M., Marumo, F., and Kimura, A. (1999). Structural analysis of the titin gene in hypertrophic cardiomyopathy:

identification of a novel disease gene. Biochemical and biophysical research communications 262, 411-417.
223. Carmignac, V., Salih, M.A., Quijano-Roy, S., Marchand, S., Al Rayess, M.M., Mukhtar, M.M., Urtizberea, J.A., Labeit, S., Guicheney, P., Leturcq, F., et al. (2007). C-terminal titin deletions cause a novel early-onset myopathy with fatal cardiomyopathy. Annals of neurology 61, 340-351.
224. Edstrom, L., Thornell, L.E., Albo, J., Landin, S., and Samuelsson, M. (1990). Myopathy with respiratory failure and typical myofibrillar lesions. Journal of the neurological sciences 96, 211-228.
225. Lange, S., Xiang, F., Yakovenko, A., Vihola, A., Hackman, P., Rostkova, E., Kristensen, J., Brandmeier, B., Franzen, G., Hedberg, B., et al. (2005). The kinase domain of titin controls muscle gene expression and protein turnover. Science 308, 1599-1603.
226. Haravuori, H., Makela-Bengs, P., Udd, B., Partanen, J., Pulkkinen, L., Somer, H., and Peltonen, L. (1998). Assignment of the tibial muscular dystrophy locus to chromosome 2q31. American journal of human genetics 62, 620-626.
227. Hackman, P., Vihola, A., Haravuori, H., Marchand, S., Sarparanta, J., De Seze, J., Labeit, S., Witt, C., Peltonen, L., Richard, I., et al. (2002). Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. American journal of human genetics 71, 492-500.
228. Ohlsson, M., Hedberg, C., Bradvik, B., Lindberg, C., Tajsharghi, H., Danielsson, O., Melberg, A., Udd, B., Martinsson, T., and Oldfors, A. (2012). Hereditary myopathy with early respiratory failure associated with a mutation in A-band titin. Brain : a journal of neurology 135, 1682-1694.
229. Pfeffer, G., Elliott, H.R., Griffin, H., Barresi, R., Miller, J., Marsh, J., Evila, A., Vihola, A., Hackman, P., Straub, V., et al. (2012). Titin mutation segregates with hereditary myopathy with early respiratory failure. Brain : a journal of neurology 135, 1695-1713.
230. Pfeffer, G., Barresi, R., Wilson, I.J., Hardy, S.A., Griffin, H., Hudson, J., Elliott, H.R., Ramesh, A.V., Radunovic, A., Winer, J.B., et al. (2014). Titin founder mutation is a common cause of myofibrillar myopathy with early respiratory failure. Journal of neurology, neurosurgery, and psychiatry 85, 331-338.
231. Udd, B., Partanen, J., Halonen, P., Falck, B., Hakamies, L., Heikkila, H., Ingo, S., Kalimo, H., Kaariainen, H., Laulumaa, V., et al. (1993). Tibial muscular dystrophy. Late adult-onset distal myopathy in 66 Finnish patients. Archives of neurology 50, 604-608.
232. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 195-201.
233. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. Nature protocols 4, 1-13.
234. Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic acids research 35, D301-303.
235. Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic acids research 37, D387-392.
236. Schrodinger, LLC. (2010). The PyMOL Molecular Graphics System, Version 1.3r1. In. (

237. Marino, M., Zou, P., Svergun, D., Garcia, P., Edlich, C., Simon, B., Wilmanns, M., Muhle-Goll, C., and Mayans, O. (2006). The Ig doublet Z1Z2: a model system for the hybrid analysis of conformational dynamics in Ig tandems from titin. Structure 14, 1437-1447.
238. Lee, E.H., Gao, M., Pinotsis, N., Wilmanns, M., and Schulten, K. (2006). Mechanical strength of the titin Z1Z2-telethonin complex. Structure 14, 497-509.
239. Chauveau, C., Bonnemann, C.G., Julien, C., Kho, A.L., Marks, H., Talim, B., Maury, P., Arne-Bes, M.C., Uro-Coste, E., Alexandrovich, A., et al. (2014). Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. Human molecular genetics 23, 980-991.
240. Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends in biochemical sciences 23, 198-199.
241. McIntosh, I., Hamosh, A., and Dietz, H.C. (1993). Nonsense mutations and diminished mRNA levels. Nature genetics 4, 219.
242. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annual review of biophysics and biomolecular structure 29, 291-325.
243. Herman, D.S., Lam, L., Taylor, M.R., Wang, L., Teekakirikul, P., Christodoulou, D., Conner, L., DePalma, S.R., McDonough, B., Sparks, E., et al. (2012). Truncations of titin causing dilated cardiomyopathy. The New England journal of medicine 366, 619-628.
244. Guo, W., Bharmal, S.J., Esbona, K., and Greaser, M.L. (2010). Titin diversity--alternative splicing gone wild. Journal of biomedicine & biotechnology 2010, 753675.
245. Pugh, T.J., Kelly, M.A., Gowrisankar, S., Hynes, E., Seidman, M.A., Baxter, S.M., Bowser, M., Harrison, B., Aaron, D., Mahanta, L.M., et al. (2014). The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. Genetics in medicine : official journal of the American College of Medical Genetics 16, 601-608.
246. van Spaendonck-Zwarts, K.Y., Posafalvi, A., van den Berg, M.P., Hilfiker-Kleiner, D., Bollen, I.A., Sliwa, K., Alders, M., Almomani, R., van Langen, I.M., van der Meer, P., et al. (2014). Titin gene mutations are common in families with both peripartum cardiomyopathy and dilated cardiomyopathy. European heart journal 35, 2165-2173.
247. Peled, Y., Gramlich, M., Yoskovitz, G., Feinberg, M.S., Afek, A., Polak-Charcon, S., Pras, E., Sela, B.A., Konen, E., Weissbrod, O., et al. (2014). Titin mutation in familial restrictive cardiomyopathy. International journal of cardiology 171, 24-30.
248. Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature reviews Genetics 14, 681-691.
249. Biesecker, L.G., Shianna, K.V., and Mullikin, J.C. (2011). Exome sequencing: the expert view. Genome biology 12, 128.
250. Samuels, D.C., Han, L., Li, J., Quanghu, S., Clark, T.A., Shyr, Y., and Guo, Y. (2013). Finding the lost treasures in exome sequencing data. Trends in genetics : TIG 29, 593-599.
251. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews Genetics 12, 745-755.

252. Samarakoon, P.S., Sorte, H.S., Kristiansen, B.E., Skodje, T., Sheng, Y., Tjonnfjord, G.E., Stadheim, B., Stray-Pedersen, A., Rodningen, O.K., and Lyle, R. (2014). Identification of copy number variants from exome sequence data. BMC genomics 15, 661.
253. Lohmann, K., and Klein, C. (2014). Next generation sequencing and the future of genetic diagnosis. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics 11, 699-707.
254. Liu, D., Ma, C., Hong, W., Huang, L., Liu, M., Liu, H., Zeng, H., Deng, D., Xin, H., Song, J., et al. (2014). Construction and analysis of high-density linkage map using high-throughput sequencing data. PloS one 9, e98855.
255. Ring, H.Z., Kwok, P.Y., and Cotton, R.G. (2006). Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7, 969-972.
256. Cotton, R. (2014). Human variome project - current overview. Molecular cytogenetics 7, I1.
257. Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W.A., Jiang, H., and Feng, G. (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer informatics 13, 67-82.
258. Biesecker, L.G., and Green, R.C. (2014). Diagnostic clinical genome and exome sequencing. The New England journal of medicine 371, 1170.
259. Biesecker, L.G., and Green, R.C. (2014). Diagnostic clinical genome and exome sequencing. The New England journal of medicine 370, 2418-2425.
260. Newman, W.G., and Black, G.C. (2014). Delivery of a clinical genomics service. Genes 5, 1001-1017.
261. Wu, L., Schaid, D.J., Sicotte, H., Wieben, E.D., Li, H., and Petersen, G.M. (2014). Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. Journal of medical genetics.
262. Agrawal, P.B., Pierson, C.R., Joshi, M., Liu, X., Ravenscroft, G., Moghadaszadeh, B., Talabere, T., Viola, M., Swanson, L.C., Haliloglu, G., et al. (2014). SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. American journal of human genetics 95, 218-226.
263. Ong, R.W., AlSaman, A., Selcen, D., Arabshahi, A., Yau, K.S., Ravenscroft, G., Duff, R.M., Atkinson, V., Allcock, R.J., and Laing, N.G. (2014). Novel cofilin-2 (CFL2) four base pair deletion causing nemaline myopathy. Journal of neurology, neurosurgery, and psychiatry 85, 1058-1060.
264. Chang, C.J., Chen, P.L., Yang, W.S., and Chao, K.M. (2014). A fault-tolerant method for HLA typing with PacBio data. BMC bioinformatics 15, 296.
265. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one 7, e47768.
266. Okoniewski, M.J., Meienberg, J., Patrignani, A., Szabelska, A., Matyas, G., and Schlapbach, R. (2013). Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers. BioTechniques 54, 98-100.
267. Steinberg, K.M., Schneider, V.A., Graves-Lindsay, T.A., Fulton, R.S., Agarwala, R., Huddleston, J., Shiryev, S.A., Morgulis, A., Surti, U., Warren, W.C., et al. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. Genome research 24, 2066-2076.

268. Quick, J., Quinlan, A.R., and Loman, N.J. (2014). A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. GigaScience 3, 22.
269. Narum, S. (2014). 2014 NGS Field Guide – Table 3c – Error rates. Available from: http://www.molecularecologist.com/next-gen-table-3c-2014/. In. (The Molecular Ecologist).
270. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome biology 14, R51.

Appendices

Appendix A

1.1 Q-solution PCR mix (Qiagen HotStar Taq Plus)

Reagent	1x
Water	9.3µl
10x Buffer	2.5µl
Q-Solution	5µl
Primers (F+R @50ng/ µl each)	2µl
dNTPs (5mM)	1µl
DNA (10ng/µl)	5µl
HotStar Taq Plus	0.2µl
Total	25µl

1.2 Standard PCR mix (Qiagen HotStar Taq Plus)

Reagent	1x
Water	14.3µl
10x Buffer	2.5µl
Q-Solution	0µl
Primers (F+R @50ng/ µl each)	2µl
dNTPs (5mM)	1µl
DNA (10ng/µl)	5µl
HotStar Taq Plus	0.2µl
Total	25µl

1.3 KOD Polymerase PCR mix (Novagen)

Reagent	1x
Water	10.5µl
10x Buffer	2.5µl
25mM $MgSO_4$	1.5µl
Primers (F+R @50ng/ µl each)	1.5µl
DMSO	1µl
dNTPs (2mM)	2.5µl
DNA (10ng/µl)	5µl
KOD Polymerase	0.5µl
Total	25µl

1.4 BigDye Terminator v3.1 mix (ABI)

Reagent	1x
Big Dye mastermix	2µl
Sequencing buffer	2µl
Primer (Forward OR Reverse)	1µl
Template DNA	2µl
Water	3µl
Total	10µl

1.5 TD63 Thermocycling protocol

Step	Temperature. (°C)	Time
1	95	5 min
2	94	30 s
3	63, -0.5 per cycle	30 s
4	72	1 min
5	Goto 2, repeat 14x	
6	94	30 s
7	55	30 s
8	72	1 min
9	Goto 6 repeat 24x	
10	72	10 min
11	10	Hold

1.6 TD65 Thermocycling protocol

Step	Temperature. (°C)	Time
1	95	5 min
2	94	30 s
3	65, -0.5 per cycle	30 s
4	72	1 min
5	Goto 2, repeat 14x	
6	94	30 s
7	58	30 s
8	72	1 min
9	Goto 6 repeat 24x	
10	72	10 min
11	10	Hold

1.7 TD68 Thermocycling protocol

Step	Temperature. (°C)	Time
1	95	5 min
2	94	30 s
3	68, -0.5 per cycle	30 s
4	72	1 min
5	Goto 2, repeat 14x	
6	94	30 s
7	60	30 s
8	72	1 min
9	Goto 6, repeat 24x	
10	72	10 min
11	10	Hold

1.8 Slowdown70 Thermocycling protocol

Step	Temperature. (°C)	Time
1	95	5 min
2	95	45 s
3	70	30 s
4	72	1 min
5	95	45 s
6	70	30 s
7	72	1 min
8	95	45 s
9	70, -1.0 per cycle	30 s
10	72	1 min
11	Goto 2, repeat 9x	
12	95	45 s
12	58	30 s
13	72	1 min
14	Goto 12, repeat 19x	
15	72	10 min
16	10	Hold

Thermal ramping set to 2.5°C/s heating, 1.5°C/s cooling

1.9 ABI Sequencing thermocycling protocol

Step	Temperature (°C)	Time
1	96	1 min
2	96	30 s
3	50	30 s
4	60	4 min
5	Goto step 2 24 times	
6	10	Hold

2.1 KBTBD13 amplification primer sets

Primers tagged with M13F gtaaaacgacggccagt, M13R ggaaacagctatgaccatg for sequencing.

Name	Sequence (5`-3`)
KBTBD13Ex1AFM13	gtaaaacgacggccagtAGAGTTGGTGGCTGGCTAAC
KBTBD13Ex1ARM13	ggaaacagctatgaccatgAGCAACGTGCTGGCCTCCAG
KBTBD13Ex1BFM13	gtaaaacgacggccagtCTGTGTTACCTGGACGAGGA
KBTBD13Ex1BRM13	ggaaacagctatgaccatgAGTGGCCATAGCTCTGCGGA
KBTBD13Ex1CFM13	gtaaaacgacggccagtCTTGTGGCCGTCTCTTCGTG
KBTBD13Ex1CRM13	ggaaacagctatgaccatgCTCACACACTCAGCCATAGC

2.2 KBTBD13 SSCP primer set

Name	Sequence (5`-3`)
KBTBD13_SSCP_F	ACGCTGGGCAACAAGCTTTA
KBTBD13_SSCP_R	CACGAAGAGACGGCCACAAG

2.3 KBTBD13 Victorian mutant screening set

Name	Sequence (5`-3`)
KBTBD13_VICMUTF	GCAAGGAGGTGGTAGAGCTG
KBTBD13_VICMUTR	TGATGGGCGTCCTCTGGAA

2.4a KBTBD13 primers for full-length amplification and cloning into dsRED2-N1

Name	Sequence (5`-3`)
KBTBD13_XhoI_F	CTCGAGATGGCACGGGGTCCACAGA
KBTBD13_KpnI_R	GGTACCCAGTTCTGCCGTTGTCGAAG

2.4b KBTBD13 primers for propeller domain amplification and cloning into dsRED2-N1

Name	Sequence (5`-3`)
KBTBD13_XhoI_Pr_F	CTCGAGAACTGCGCATTGCTGTGC
KBTBD13_KpnI_R	GGTACCCAGTTCTGCCGTTGTCGAAG

2.5 KBTBD13 primers for full-length amplification and cloning into pETM-11

Name	Sequence (5`-3`)
KBTBD13_NcoI_F	CCATGGCACGGGGTCCACAGA
KBTBD13_KpnI_SR	GGTACCTCACAGTTCTGCCGTTGTCG

2.6 KBTBD13 primers for full-length amplification and cloning into pET-44a

Name	Sequence (5`-3`)
KBTBD13_BamH1_F	GGATCCATGGCACGGGGTCCACAGA
KBTBD13_KpnI_R2	GGTACCGTCAGTTCTGCCGTTGTCGAAG

2.7 Primers for amplifying *TMEM33* and *ZFHX4* variants

Name	Sequence (5`-3`)
TMEM33 Ex3F	GATGCAATTCACAAAGGAGG
TMEM33 Ex3R	ACAAGAGGCAAGAACAAAGATAG
ZFHX4 V1 F2	CTGTGCATGATCATCGGAT
ZFHX4 V1 R2	GGCAAAGAGTCACCATTTTC
ZFHX4 V2 F2	AACCGCTCTCTGTTTCTGAC
ZFHX4 V2 R2	CCCACAGTTCTCCATTCAC

2.8 TTN mutant screening primer set

Name	Sequence (5`-3`)
TTN Ex3 F	TCGGACAAGGCAGTGAAC
TTN Ex3 R	TTGCTGGAGATGTCCTCTG
TTN Ex95 F	TTAGCAATGAAGCTCCGAAG
TTN Ex95 R	ACATGCCAGATCATCGATTG
TTN Ex343 F	GAAAGAATGATCTTCATCATGG
TTN Ex343 R	CCCCCTGATTCTATTACATTTC

Appendix B

Table of genes in the NSES panel, arranged in alphabetical order according to disease category. Duplicate entries are present when a disease gene is associated with multiple disease phenotypes.

Ataxias	Cardiac disease		Congenital muscular dystrophy	CMS	CMT	
ADCK3	ABCC9	MYOZ2	B3GALNT2	AGRN	AARS	MYH7
AFG3L2	ACTC1	MYPN	CAPN3	CHAT	AIFM1	NDRG1
ANO10	ACTN2	NEXN	CAV3	CHRNA1	ARHGEF10	NEFL
APTX	ANK2	NPPA	COL6A1	CHRNB1	ATL1	NGF
ATM	ANKRD1	PKP2	COL6A2	CHRND	ATP7A	NTRK1
ATP2B3	BAG3	PLN	COL6A3	CHRNE	BSCL2	PDK3
BEAN1	CACNA1C	PRKAG2	DAG1	CHRNG	CCT5	PLEKHG5
C10orf2	CACNB2	PSEN1	DMD	COLQ	CTDP1	PMP22
CACNA1A	CASQ2	PSEN2	DNM2	DOK7	DCTN1	PRPS1
CACNB4	CAV3	RBM20	DPM2	DPAGT1	DHKTD1	PRX
FGF14	COX15	RYR2	DYSF	GFPT1	DNM2	RAB7A
FXN	CSRP3	SCN5A	EMD	LAMB2	DNMT1	SBF2
IFRD1	DES	SDHA	FHL1	MUSK	DYNC1H1	SEPT9
ITPR1	DMD	SGCD	FKRP	PLEC	EGR2	SETX
KCNA1	DSC2	SLC25A4	FKTN	RAPSN	FAM134B	SH3TC2
KCNC3	DSG2	TAZ	GMPPB	SCN4A	FBLN5	SLC12A6
MRE11A	DSP	TCAP	GTDC2		FGD4	SOX10
MTTP	DTNA	TGFB3	ISPD		FIG4	SPTLC1
PEX7	EYA4	TMEM43	ITGA7		GAN	SPTLC2
PHYH	FKTN	TMPO	LAMA2		GARS	TFG
POLG	GATAD1	TNNC1	LARGE		GDAP1	TRPV4
PRKCG	GJA5	TNNI3	LMNA		GJB1	WNK1
SACS	GPD1L	TNNT2	MYOT		GJB3	YARS
SETX	HCN4	TPM1	POMGNT1		HARS	
SIL1	JPH2	TTN	POMT1		HINT1	
SLC1A3	JUP	TTR	POMT2		HK1	
SPTBN2	KCNA5	VCL	SEPN1		HOXD10	
SYNE1	KCNE1		SGCA		HSPB1	
TDP1	KCNE2		SGCB		HSPB8	
TK2	KCNE3		SGCD		IFRD1	
TTBK2	KCNH2		SGCG		IGHMBP2	
TTPA	KCNJ2		TCAP		IKBKAP	
	KCNQ1		TRIM32		INF2	
	LAMA4		TTN		KARS	
	LAMP2				KIF1A	
	LDB3				KIF1B	
	LMNA				LITAF	
	MYBPC3				LMNA	
	MYH6				LRSAM1	
	MYH7				MED25	
	MYL2				MFN2	
	MYL3				MPZ	
	MYLK2				MTMR2	

Distal arthrogryposis		Distal myopathy	Glycogen storage disease/Rhabdomyolysis	Inclusion body myositis
ACTA1	MPZ	ANO5	ACADL	ANO5
BIN1	MTM1	CAV3	ACADVL	DYSF
CHAT	MUSK	CRYAB	AGL	GNE
CHRNA1	MYBPC1	DES	CPT1B	LDB3
CHRNB1	MYH2	DNM2	CPT2	MYH7
CHRND	MYH3	DYSF	ENO3	MYOT
CHRNE	MYH8	FHL1	GAA	TTN
CHRNG	NEB	FLNC	GBE1	VCP
CNTN1	PAFAH1B1	GNE	GYG1	
COL6A1	PEX7	KLHL9	GYS1	
COL6A2	PFKM	LDB3	LDHA	
COL6A3	PIP5K1C	MATR3	LPIN1	
DOK7	PMP22	MYH7	PFKM	
EGR2	POMGNT1	MYOT	PGAM2	
ERBB3	POMT1	NEB	PGK1	
FGFR2	POMT2	TCAP	PGM1	
FHL1	PRX	TIA1	PHKA1	
FKRP	RAPSN	TTN	PYGM	
FKTN	RELN	VCP	RYR1	
FLNA	RYR1			
GBE1	SEPN1			
GLE1	SOX10			
GTDC2	SYNE1			
HSPG2	TNNI2			
ISPD	TNNT3			
KCNA1	TPM2			
L1CAM	TRPV4			
LAMA2	UBA1			
LARGE	UTRN			
LMNA				

Hereditary spastic paraplegia		Lissencephaly	Limb-girdle muscular dystrophy		Miscellaneous	
ABCD1	SLC33A1	ARX	ACADVL	MYOT	ABCD1	PHYH
ALDH3A2	SPAST	DCX	ANO5	PABPN1	ABHD12	PRRT2
ALS2	SPG11	PAFAH1B1	CAPN3	PLEC	AHNAK	PTRF
AP5Z1	SPG20	RELN	CAV3	POMGNT1	ARVC1	SGCE
ATL1	SPG21	TUBA1A	CPT2	POMT1	APOA1	SCL22A5
BSCL2	SPG7	TUBB3	DAG1	POMT2	ARSA	SCL25A20
CYP7B1	ZFYVE26		DES	SEPN1	DOCK3	SOX10
FA2H	ZFYVE27		DNAJB6	SGCA	ETFA	TOR1A
HSPD1			DYSF	SGCB	ETFB	
KIAA0196			FKRP	SGCD	ETFSH	
KIF1A			FKTN	SGCG	FA2H	
KIF5A			GMPPB	SMCHD1	ILK	
L1CAM			GTDC2	SYNE2	KIF21A	
NIPA1			ISPD	TCAP	KLHL9	
PLP1			LAMA2	TNPO3	MSTN	
PNPLA6			LARGE	TRIM32	MURC	
REEP1			LMNA	TTN	NOTCH3	
SACS			MEGF10		PHOX2A	

Mitochondrial	Myofibrillar myopathy	Motor neuron disease	Myopathies	Myotonia/ Channelopathies
MRPL3	BAG3	AARS	ABHD5	CACNA1S
NDUFAF1	CRYAB	ALS2	ACTA1	CLCN1
OPA1	DES	ANG	ATP2A1	CNBP
POLG	DNAJB6	AR	BIN1	KCNJ18
POLG2	FHL1	ASAH1	CFL2	SCN4A
RRM2B	FLNC	ATP7A	CNTN1	
SLC25A4	LDB3	BSCL2	COL6A1	
SUCLA2	MYOT	DCTN1	COL6A2	
TK2	PLEC	DNAJB2	COL6A3	
	VCP	DYNC1H1	DNM2	
		ERBB3	FHL1	
		FIG4	GMPPB	
		FUS	ISCU	
		GARS	KBTBD13	
		GLE1	KLHL40	
		HSPB1	KLHL41	
		HSPB3	LAMP2	
		HSPB8	LMOD3	
		IGHMBP2	MATR3	
		PFN1	MEGF10	
		PIP5K1C	MTM1	
		PLEKHG5	MYBPC3	
		REEP1	MYH2	
		SEPT9	MYH7	
		SETX	MYL2	
		SOD1	NEB	
		TARDBP	PNPLA2	
		TRPV4	RYR1	
		VAPB	SEPN1	
		VCP	STIM1	
		VRK1	TNNT1	
			TPM2	
			TPM3	
			TRIM32	
			TTN	
			VMA21	