Identifying and modelling genes that are associated with rare developmental disorders

Keren Jacqueline Carss
Wellcome Trust Sanger Institute
Queens' College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
August 2014

Declaration

I declare that this dissertation describes my own, original work. It only includes work done in collaboration where clearly and specifically indicated in the text. No part of this dissertation has been submitted for any other degree, diploma, or qualification at any university or institution. This dissertation does not exceed 60,000 words.

Keren Jacqueline Carss

August 2014

Abstract

While individually rare, collectively developmental disorders are common, affecting around 3% of live births in the UK. The aetiology of these disorders often includes a genetic component. Next generation sequencing provides a powerful tool with which to identify variants that cause rare developmental disorders. This dissertation describes three distinct projects in which next generation sequencing was used for this purpose, along with statistical or functional follow-up approaches.

A cohort of 30 fetuses with a diverse range of structural abnormalities, along with their parents, was exome sequenced. I analysed these data to identify rare, high quality, coding variants consistent with a *de novo* or recessive inheritance model. I investigated several methods of variant interpretation, including manual and computational methods, and found causative variants for 10% of the cohort. These results suggest that next generation sequencing is a promising method for prenatal genetic diagnostics.

As part of the UK10K project, 996 patients with moderate to severe intellectual disability (ID) underwent sequencing of 565 known or candidate ID-associated genes. I developed and implemented a pipeline to identify likely causative loss of function (LOF) variants through extensive quality filtering. From these data, causative variants were identified for ~14% of the cohort, and the novel ID-associated gene *SETD5* was identified. Next, I performed a series of case-control enrichment analyses to evaluate the contribution of different classes of possibly pathogenic variants. Patients with ID had a significant enrichment of both LOF and missense variants in known ID-associated genes, compared to controls with non-syndromic congenital heart defects.

One strategy to investigate the consequences of a potentially pathogenic variant is to inhibit expression of the gene in an appropriate animal model, and assess the extent to which aspects of the human phenotype are recapitulated. I applied this technique to two genes identified from the UK10K project as likely to be associated with dystroglycanopathy, a subtype of muscular dystrophy. I inhibited the expression of both genes, *B3GALNT2* and *GMPPB*, in zebrafish embryos using morpholino oligonucleotides. The phenotype of both models mimicked several aspects of the human phenotype including morphological defects such as micropthalmia and hydrocephalus, structural defects of the tissue such as disordered muscle fibres, and the precise molecular defect, which is hypoglycosylation of α-dystroglycan.

Acknowledgements

Throughout my PhD, I have been involved in several large collaborative projects, so there are a great many people who I would like to thank. My first and biggest thank you is to my primary supervisor, Dr Matthew Hurles. He provided me with the opportunity to work on several exciting and fruitful projects. He also provided the perfect balance of support and freedom, along with a constant stream of outstanding ideas, and encouragement when I needed it. Thank you also to my secondary supervisor, Dr Derek Stemple. He introduced me to the zebrafish, and gave me valuable support throughout my time working in the zebrafish laboratory. Thank you to the additional members of my thesis committee: Dr Helen Firth and Dr David Adams, for guidance and advice throughout the last four years.

Thank you to the following clinical collaborators at the University of Birmingham: Prof Eamonn Maher, Prof Mark Kilby, Dr Dominic McMullan, and Dr Sarah Hillman. I very much enjoyed working on the abnormal fetal development project, and I learned a lot from it. Thank you to Dr Vijaya Parthiban, Dr Alejandro Sifrim and Dr Damian Smedley for running the CoNVex, eXtasy, and PhenoDigm programs respectively during this project.

I am grateful to the UK10K consortium, particularly the rare disease group, for the opportunity to be involved in two exciting UK10K projects. Thank you to Dr Lucy Raymond for the opportunity to work on the UK10K intellectual disability cohort, in an exciting and fruitful collaboration. I also thank the many other people involved in this project, most importantly Dr Detelina Grozeva, Dr Olivera Spasic-Boskovic, and Dr James Floyd.

Thank you to Prof Francesco Muntoni for giving me the opportunity to work on the UK10K dystroglycanopathy project. Also to other members of this project including Dr Elizabeth Stevens, Dr Silvia Torelli, Dr Sebahattin Cirak, and Dr Reghan Foley. I owe a particularly large thank you to Dr Yung-Yao Lin, who provided a huge amount of support, supervision, and training to me during this project. Thank you also to Dr Sebastian Gerety for help and support with designing zebrafish experiments.

Thank you to all patients who participate in research studies, without whose generosity none of this work would have been possible.

I am very grateful to the Wellcome Trust for generously funding my PhD. I have been fortunate to do my PhD at the Wellcome Trust Sanger Institute, which provides not only

a very stimulating environment, but also unparalleled facilities. I therefore thank the many people who work in the various pipelines and services teams at the Wellcome Trust Sanger Institute, including the sample reception and sequencing teams, the GAPI team, Human Genome Informatics, and the many other people who keep the laboratory and computational facilities running. I am grateful to Dr Alex Bateman, Dr Julian Rayner, Dr Annabel Smith, Christina Hedberg-Delouka, and Carol Dunbar for practical and administrative support.

I thank all past and present members of teams 29 and 31 for help with experiments or coding, and for many fun and enlightening conversations over the years. I would like to especially thank Dr Saeed Al Turki, Dan King, and Dr Vijaya Parthiban for their great generosity in sharing their scripts, skills, and time with me. I am also particularly grateful to Dr Ana Cvejic and Dr Jovana Serbanovic-Canic for teaching me basic experimental techniques in zebrafish when I first started.

On a more personal note, I thank the many additional scientists who have encouraged, mentored, and inspired me throughout my career, both directly and by example. These include Prof Graham Anderson, Dr Kevin O'Shaughnessy, Dr Annie Mercier Zuber, Dr Darren Logan, Dr Inês Barrosso, Dr Eleanor Wheeler, Dr Eleftheria Zeggini, Prof Jane Worthington, and Prof Eamonn Maher. Last but not least I thank my friends and family, especially the fabulous PhD10, the wonderful Norman and Carss families, and my endlessly supportive husband.

Publications

Publications arising from work associated with this thesis:

- Mackie FL, **Carss KJ**, Hillman SC, Hurles ME, & Kilby MD. Exome sequencing in fetuses with structural malformations. [Review article] ***Journal of Clinical Medicine.*** 2014 July, 3(3), 747-762.

- Grozeva D, **Carss KJ**, Spasic-Broskovic O, Parker M, Archer H, Firth HV, *et al. De novo* mutations in *SETD5* cause intellectual disability and associated features of 3p25 microdeletion syndrome. ***American Journal of Human Genetics****.* 2014 April, 3;94(4):618-24.

- **Carss KJ**, Hillman SC, Parthiban V, McMullan DJ, Maher ER, Kilby MD & Hurles ME. Exome sequencing improves genetic diagnosis of structural fetal abnormalities revealed by ultrasound. ***Human Molecular Genetics.*** 2014 June, 15,23(12):3269-77.

- **Carss KJ***, Stevens E*, Foley AR, Cirak S, Riemersma M, Torelli S, *et al*. Mutations in *GDP-Mannose pyrophosphorylase B* cause congenital and limb-girdle muscular dystrophies associated with hypoglycosylation of α-dystroglycan. ***American Journal of Human Genetics.*** 2013 July, 11;93(1):29-41.

- Stevens E*, **Carss KJ***, Cirak S, Foley AR, Torelli S, Willer T, *et al.* Mutations in *B3GALNT2* cause congenital muscular dystrophy and hypoglycosylation of α-dystroglycan. ***American Journal of Human Genetics.*** 2013 March, 7;92(3):354-65.

* Jointly contributing authors.

Table of Contents

1 Introduction 1

2 Exome sequencing improves genetic diagnosis of structural fetal abnormalities 5

2.1 Introduction 5

2.1.1 The impact and causes of fetal structural abnormalities 5

2.1.2 Current techniques for prenatal genetic diagnosis 6

2.1.3 Next generation sequencing 9

2.1.4 Variant prioritisation strategies 10

2.1.5 Prenatal next generation sequencing: proof of concept 12

2.1.6 Aims, context, and colleagues 12

2.2 Methods 14

2.2.1 Cohort 14

2.2.2 Exome sequencing 15

2.2.3 VCF file merging, annotation, and quality control 15

2.2.4 Identification of *de novo* SNVs and indels 16

2.2.5 Identification of inherited recessive and X-linked SNVs and indels 17

2.2.6 Identification of CNVs 17

2.2.7 Sanger sequencing 18

2.2.8 Interpretation of variants 18

2.3 Results 21

2.3.1 The exome sequencing data are of high quality 21

2.3.2 There is a mean of 1.13 validated *de novo* SNVs or indels per fetus 26

2.3.3 There are three candidate *de novo* or X-linked copy number variants 28

2.3.4 There is a mean of 13 candidate genes with inherited recessive or X-linked variants per fetus, in the preliminary round of analysis 30

2.3.5 *De novo* SNVs in *FGFR3* and *COL2A1* are highly likely to be causal 30

2.3.6 *De novo* SNVs in *NF1* and *SMARCC2* are possibly causal 32

2.3.7 Two unrelated fetuses with no clear clinical overlap have *de novo* SNVs in *PARD3B* 33

2.3.8 A *de novo* deletion that overlaps with *OFD1* is highly likely to be causal 34

2.3.9 Inherited recessive or X-linked SNVs in five fetuses are possibly causal, in the preliminary round of analysis 34

2.3.10 The variant prioritisation program eXtasy identifies 36 possibly causal variants, with an enrichment of *de novo* mutations 36

2.3.11 The variant prioritisation program PhenoDigm identifies possibly causal variants in 18 genes ... 39

2.3.12 There is a degree of overlap between the variants identified as possibly causal by the three different prioritisation methods ... 41

2.3.13 The continuing need for manual curation ... 42

2.3.14 Inherited recessive or X-linked SNVs in five fetuses are possibly causal, in the final round of analysis ... 42

2.3.15 The estimated diagnostic yield of this study is 10% ... 44

2.4 Discussion ... 47

2.4.1 Summary ... 47

2.4.2 The diagnostic yield in context ... 47

2.4.3 Comparison of variant interpretation methods ... 48

2.4.4 The ethics of next generation sequencing for prenatal genetic diagnosis ... 50

2.4.5 Next generation sequencing is the future of prenatal genetic diagnostics 52

3 Case-control analysis of 565 known and candidate intellectual disability-associated genes ... 55

3.1 Introduction ... 55

3.1.1 The impact of intellectual disability ... 55

3.1.2 Discovery of intellectual disability-associated genes ... 57

3.1.3 Biology of intellectual disability-associated genes ... 59

3.1.4 Case-control enrichment analysis of rare variants ... 62

3.1.5 Aims, context, and colleagues ... 65

3.2 Methods ... 67

3.2.1 Samples, sequencing, and quality control ... 67

3.2.2 Annotation and filtering pipeline ... 67

3.2.3 Principal component analysis ... 68

3.2.4 Cohort allelic sums test ... 69

3.3 Results ... 72

3.3.1 Targeted resequencing of 565 intellectual disability-associated genes in cases and controls ... 72

3.3.2 The sequencing data are of good quality ... 72

3.3.3 There is no substantial difference in population structure between the intellectual disability and congenital heart disease cohorts ... 74

3.3.4 14% of intellectual disability patients have a likely causative variant in a sequenced intellectual disability-associated gene ... 76

3.3.5 *SETD5* is a novel intellectual disability-associated gene ... 77

3.3.6 Individuals with intellectual disability have an enrichment of loss of function variants in sequenced ID-associated genes, compared to controls ... 78

3.3.7 In known ID-associated genes on the X chromosome, unique missense variants tend to be more damaging in ID patients than controls. 81

3.3.8 Evidence for an enrichment of unique, predicted damaging, missense variants in sequenced ID-associated genes in the ID cohort 83

3.4 Discussion 86

3.4.1 Summary 86

3.4.2 Loss-of-function of the histone methyltransferase gene *SETD5* is probably responsible for the cardinal features of 3p25 microdeletion syndrome 86

3.4.3 Insights from case-control enrichment analyses 89

3.4.4 Limitations of this study 91

3.4.5 Further work 92

4 Modelling dystroglycanopathy in zebrafish embryos by knockdown of *B3GALNT2* and *GMPPB* 94

4.1 Introduction 94

4.1.1 The phenotypic spectrum of dystroglycanopathy 94

4.1.2 Dystroglycan structure, function and glycosylation 95

4.1.3 Known dystroglycanopathy-associated genes 96

4.1.4 Frequency of variants, and genotype-phenotype correlations 101

4.1.5 Zebrafish models of genetic disease 102

4.1.6 Animal models of dystroglycanopathy 104

4.1.7 Aims, context, and colleagues 105

4.2 Materials and methods 107

4.2.1 Sequencing of clones 107

4.2.2 Reverse transcription polymerase chain reaction 107

4.2.3 Design and injection of morpholino oligonucleotides 107

4.2.4 Generation of green fluorescent protein-tagged RNA 108

4.2.5 Immunofluorescence staining 109

4.2.6 Evans blue dye assay 109

4.2.7 Immunoblotting 109

4.3 Results 110

4.3.1 *B3GALNT2* and *GMPPB* are conserved with their zebrafish orthologues 110

4.3.2 Expression of *b3galnt2* and *gmppb* throughout early zebrafish development 114

4.3.3 Finding the optimal morpholino oligonucleotide and dose for *b3galnt2* and *gmppb* 115

4.3.4 Morpholino oligonucleotides reduce the expression of *b3galnt2* and *gmppb* 116

4.3.5 *b3galnt2* morphants have gross morphological defects including hydrocephalus and impaired motility 117

4.3.6 *b3galnt2* morphants have muscle defects including gaps in the myosepta and lesions between fibres 119

4.3.7 *b3galnt2* morphants have hypoglycosylated α-dystroglycan 120

4.3.8 Coinjection with wildtype RNA fails to rescue the *b3galnt2* morphant phenotype 120

4.3.9 *gmppb* morphants have gross morphological defects including micropthalmia and impaired motility 122

4.3.10 *gmppb* morphants have muscle defects including disordered fibres, incomplete myosepta, and interfibre spaces 124

4.3.11 *gmppb* morphants have hypoglycosylated α-dystroglycan 124

4.4 Discussion 126

4.4.1 Summary 126

4.4.2 Phenotypic rescue 126

4.4.3 The function of B3GALNT2 127

4.4.4 The function of GMPPB 128

4.4.5 Zebrafish phenotypes in context 130

4.4.6 Technical limitations of this study 133

4.4.7 Future research 133

5 Discussion 136

6 References 140

7 Appendices 179

List of Figures

Figure 2 - 1: Matrix showing categories of phenotypes in the cohort of fetuses with structural abnormalities 15

Figure 2 - 2: Decision tree for classifying candidate genes into three categories......... 20

Figure 2 - 3: Target coverage of exome sequencing reads by sample. 21

Figure 2 - 4: Quality control metrics for single nucleotide variants......... 24

Figure 2 - 5: Quality control metrics for indels. 25

Figure 2 - 6: Log_2 ratios of candidate CNVs in fetuses with structural abnormalities. .. 29

Figure 2 - 7: Pedigree of trio 23, showing Sanger sequencing of *de novo* mutation in *FGFR3*. 31

Figure 2 - 8: Pedigree of trio 20, showing Sanger sequencing of *de novo* mutation in *COL2A1*. 32

Figure 2 - 9: Venn diagram showing overlap between the genes prioritised by each of the three methods. 41

Figure 3 - 1: Quality control metrics for the UK10K targeted resequencing study........ 73

Figure 3 - 2: Principal component analysis........ 75

Figure 3 - 3: Classes of variant identified through the R filtering pipeline. 76

Figure 3 - 4: Facial appearance of individuals with *SETD5* mutations. 78

Figure 3 - 5: Patients with intellectual disability have an enrichment of loss of function variants in sequenced intellectual disability-associated genes compared to controls. 79

Figure 3 - 6: In known ID-associated genes on the X chromosome, unique missense variants are predicted to be more damaging in ID patients than controls............ 83

Figure 3 - 7: In known ID-associated genes on the autosomes, unique missense variants are not predicted to be more damaging in ID patients than controls....... 83

Figure 4 - 1: Model of α-DG interactions. 96

Figure 4 - 2: The function of GMPPB in glycosylation pathways. 106

Figure 4 - 3: Protein alignment showing conservation of B3GALNT2. 112

Figure 4 - 4: Protein alignment showing conservation of GMPPB. 113

Figure 4 - 5: Reverse transcription PCR shows expression of *b3galnt2* and *gmppb* throughout early zebrafish development. 114

Figure 4 - 6: Coinjection of *p53* MO rescues the neurodegeneration induced by *b3galnt2* MO. 115

Figure 4 - 7: The *b3galnt2* MO inhibits expression of recombinant GFP-tagged *b3galnt2* RNA. 116

Figure 4 - 8: The *gmppb* splice blocking MO disrupts RNA splicing. 117

Figure 4 - 9: *b3galnt2* knockdown zebrafish embryos have muscle defects and hypoglycosylated α-DG at 48 hpf. 118

Figure 4 - 10: Coinjection with wildtype human *B3GALNT2* RNA fails to rescue the *b3galnt2* morphant phenotype. 121

Figure 4 - 11: Coinjection with wildtype zebrafish *b3galnt2* RNA fails to rescue the *b3galnt2* morphant phenotype. 122

Figure 4 - 12: *gmppb* knockdown zebrafish embryos have morphological defects, damaged muscle, and hypoglycosylated α-DG at 48 hpf. 123

Figure 4 - 13: α-DG of *gmppb* morphants is hypoglycosylated relative to wildtype embryos. 125

Figure 4 - 14: B3GALNT2 catalyses the synthesis of the trisaccharide GalNAcβ1-3-GlcNAc-β1,4-Man, which is required for laminin binding. 128

List of Tables

Table 2 - 1: Exome sequencing coverage and quality control metrics. 23

Table 2 - 2: Validated *de novo* SNVs in fetuses with structural abnormalities. 27

Table 2 - 3: Candidate CNVs in fetuses with structural abnormalities. 28

Table 2 - 4: Candidate genes identified as possible causal by eXtasy. 38

Table 2 - 5: Candidate genes identified as possibly causal by PhenoDigm. 40

Table 2 - 6: Summary of all candidate genes identified in 30 fetuses with structural abnormalities. 46

Table 3 - 1: Functional classes of ID-associated genes. 59

Table 3 - 2: List of 204 sequenced intellectual disability-associated genes that are known. 70

Table 3 - 3: List of 361 sequenced intellectual disability-associated genes that are candidates. 71

Table 3 - 4: Candidate genes with the highest number of LOF variants. 77

Table 3 - 5: Enrichment of unique LOF variants in the ID cohort, split by category. 80

Table 3 - 6: Enrichment of unique, predicted damaging, missense variants in the ID cohort, split by category. 84

Table 4 - 1: Primers used to analyse *b3galnt2* and *gmppb* expression in early zebrafish development, and splicing disruption in *gmppb* morphants. 107

Table 4 - 2: Morpholino oligonucleotide sequences and predicted effects. 108

Table 4 - 3: Percentage identity of B3GALNT2 orthologues of five diverse eukaryotic species with human B3GALNT2. 111

Table 4 - 4: Percentage identity of GMPPB orthologues of five diverse eukaryotic species with human GMPPB. 111

Table 4 - 5: b3galnt2 morphants are significantly more likely to have hydrocephalus than wildtype embryos. 119

Table 4 - 6: Phenotypic comparison of zebrafish dystroglycanopathy models. 132

List of Appendices

Appendix 1: Primer sequences for Sanger sequencing of variants that passed validation. 180

Appendix 2: Inherited recessive and X-linked SNPs and indels that pass filters, in fetuses with structural abnormalities (preliminary round of analysis). 188

Appendix 3: High-quality, rare, coding, inherited recessive and X-linked SNPs and indels (final round of analysis). 192

List of Abbreviations

aCGH	Array comparative genomic hybridisation
α-DG	α-dystroglycan
ASD	Autism spectrum disorder
β-DG	β-dystroglycan
BioGPS	Biology Gene Portal System
bp	Base pair
CAST	Cohort allelic sums test
CDG	Congenital disorder of glycosylation
cfDNA	Cell-free DNA
CHD	Congenital heart disease
CK	Creatine kinase
CMD	Congenital muscular dystrophy
CNS	Central nervous system
CNV	Copy number variant
CRISPR	Clustered regularly interspaced short palindromic repeat
DDD	Deciphering developmental disorders
DDG2P	Developmental Disorder Gene2Phenotype
DGC	Dystrophin-glycoprotein complex
DNA	Deoxyribonucleic Acid
Dol-P-Man	Dolichol phosphate mannose
EBD	Evans blue dye
ECM	Extracellular matrix
ENU	*N*-ethyl-*N*-nitrosourea
ER	Endoplasmic reticulum
ESP	Exome Sequencing Project
FCMD	Fukuyama-type CMD
FISH	Fluorescence in situ hybridisation
FORGE	Finding of rare disease genes
GalNAc	N-acetylgalactosamine
Gb	Gigabases
GDP	Guanosine diphosphate
GFP	Green fluorescent protein
GlcNAc	N-acetylglucosamine
GPI	Glycosylphosphatidylinositol
GWAS	Genome-wide association study
HDL-C	High-density lipoprotein cholesterol
HMQ	High mapping quality
Hpf	Hours post fertilisation
HPO	Human phenotype ontology
IC	Information Content
ID	Intellectual disability

IEM	Inborn error of metabolism
IGV	Integrative Genomics Viewer
IKMC	International knockout mouse consortium
Indel	Insertion deletion
IQ	Intelligence quotient
Kb	Kilobase
LD	Linkage disequilibrium
LGMD	Limb girdle muscular dystrophy
LOF	Loss of function
Mb	Megabase
MEB	Muscle-eye-brain disease
MO	Morpholino oligonucleotide
MPO	Mammalian phenotype ontology
MRI	Magnetic resonance imaging
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
NHLBI EVS	National Institute of Health: Heart, Lung, and Blood Institute exome variant server
NIPT	Non-invasive prenatal testing
OMIM	Online Mendelian Inheritance in Man
PAGE	Prenatal Assessment of Genomes and Exomes
PCA	Principal component analysis
PCR	Polymerase chain reaction
PKU	Phenylketonuria
PSD	Postsynaptic density
PTR	Primary target region
QF-PCR	Quantitative fluorescent polymerase chain reaction
RD	Retinal dystrophy
RNA	Ribonucleic Acid
RT-PCR	Reverse transcription polymerase chain reaction
SB	Splice blocking
SimJ	Jaccard Index
SKAT	Sequence kernel association test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TALEN	Transcription activator-like effector nuclease
TB	Translation blocking
UTR	Untranslated region
VCF	Variant call format
VEP	Variant effect predictor
VOUS	Variant of unknown significance
WTSI	Wellcome Trust Sanger Institute
WWS	Walker-Warburg Syndrome
ZFIN	Zebrafish information network
ZFN	Zinc finger nuclease

1 Introduction

Rare genetic disorders often have a classical Mendelian pattern of inheritance, and they are often caused by a single high-penetrance variant. There are at least 6000-7000 rare genetic disorders, meaning that collectively they are in fact common, and the causes of around half have been identified thus far (1). While numerous different phenotypes are associated with rare genetic disorders, they often affect development, and first manifest *in utero*, in infancy, or in childhood.

There are two reasons why the study of rare developmental disorders is of great importance. First, it directly improves the lives of patients and their families. Occasionally, identification of the genetic cause of a disorder will lead to improved treatment or a new therapy for a patient (2). It also often allows patients and their families to access additional social and educational services, and it can allow estimation of recurrence risk for future pregnancies. Families affected by a rare developmental disorder often go through a ‘diagnostic odyssey’ that can last a decade or more, during which many different individual medical and genetic tests are performed in an attempt to identify the cause of the disorder (3). Therefore, finally receiving a genetic diagnosis can bring great peace of mind, even if it would not influence treatment.

The second reason to study rare developmental disorders is that they often give insights into relevant biological processes, and into the aetiology of more common forms of disease. This has been recognised for centuries. In 1657 Dr William Harvey observed that “there is no better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of nature by the careful investigation of cases of rarer forms of disease.” For example, pathogenic variants in *PFN1* can cause familial amyotrophic lateral sclerosis (ALS), and they have also been implicated in the sporadic form of the disorder (4). Furthermore, this finding suggested that dysregulation of cytoskeletal machinery has role in the aetiology of ALS. In another example, pathogenic variants in several member of the SWI/SNF complex, which is involved in chromatin remodelling, can cause Coffin-Siris syndrome, highlighting the importance of appropriate chromatin remodelling (5).

Historically, identification of genes associated with rare developmental disorders relied on linkage mapping followed by positional cloning or painstaking Sanger sequencing of candidate genes. Many genes were identified in this way, including *CFTR* in cystic fibrosis, to name but one example (6). However, this method requires large families with multiple affected individuals, a relatively homogeneous and high-penetrance phenotype, and often knowledge of the function of candidate genes, which severely limits the utility of this approach. However, in recent years, the development of next generation sequencing (NGS) has enabled the entire genome (or selected portions of it such as the exome) to be sequenced in a rapid, systematic, high-throughput, and relatively cheap manner. This has led to nothing less than a revolution in the field of rare developmental disorder diagnostics and gene discovery.

The first example of NGS to identify a novel rare disorder-associated gene came in 2010, when pathogenic variants in *DHODH* were found to cause Miller syndrome (7). Since then, at least one hundred other rare disorder-associated genes have been identified through the application of NGS, bringing many advantages both directly to the lives of those patients, and indirectly to the wider understanding of the pathogenesis of developmental disorders (8). Several consortia around the world have been established to sequence the exomes or genomes of cohorts of patients with rare genetic disorders on a large scale, including the Deciphering Developmental Disorders (DDD) project, the UK10K project, the Finding of Rare Disease Genes (FORGE) Canada Consortium and others (3, 9-11).

Recently, there has been much discussion surrounding the exact extent and nature of the evidence required in order to state that a given gene is indeed associated with a rare genetic disorder. While there are still contentions in this area, the importance of a consistent and stringent approach is increasingly being recognised, and a preliminary set of guidelines for this purpose was recently published (12). Identification of a loss of function variant that segregates with a rare disorder in a single family is not on its own sufficient evidence that the variant causes that disorder, particularly because loss of function variants in many genes are not uncommon in healthy individuals (12, 13). Therefore, statistical or functional follow-up experiments are also required.

One very important and commonly used statistical follow-up approach is to identify potentially pathogenic variants in the same gene in multiple unrelated affected individuals (3). There is no one rule as to the number of unrelated individuals required to statistically demonstrate that the occurrence of a particular number of variants in a particular gene is highly unlikely to occur by chance. Instead, the number required depends on various factors including the size of the gene, and its mutation rate. Another relevant statistical follow-up approach that can be used is identification of a significant burden of variants in cases compared to controls (14).

Functional follow-up approaches can be an alternative or complementary method to statistical follow-up approaches. Examples of such approaches include *in silico* experiments such as computational modelling of the effect of a variant on the structure of a protein (15), *in vitro* experiments such as investigation of the affect of a variant in human cells (16), and *in vivo* experiments such as recapitulation of aspects of patients' phenotypes using an appropriate animal model (17). Selection of appropriate statistical or functional follow-up experiments for the study of putative rare disorder-associated genes is of great importance, and depends on many factors including the availability of additional patients with overlapping phenotypes, the mode of inheritance of the phenotype, the predicted mechanism of action of the variant, and current knowledge of gene function.

In this dissertation I describe three distinct projects in which NGS was used to identify variants that cause rare developmental disorders, followed by statistical or functional follow-up approaches to validate or further explore the results. Because the projects are distinct, the following three chapters are self-contained, and the majority of the introductory and discursive material is located within each chapter.

The aim of the project described in chapter 2 was to explore how well exome sequencing performs as a method for identifying variants that cause abnormal fetal development, by performing exome sequencing on 30 parent-fetus trios where the fetuses had a diverse range of structural abnormalities. In chapter 2 I will describe the analysis of these data, different methods of interpreting variants, and the identification of causal and possibly causal variants. This project demonstrates that exome sequencing is a promising method for prenatal genetic diagnostics.

In chapter 3 I will describe a targeted resequencing study that was performed on a cohort of patients with intellectual disability (ID) as part of the UK10K project. The aims of this project were to identify causal variants in known ID-associated genes in the cohort, to identify novel ID-associated genes, and to ascertain whether there is a burden of variants in ID-associated genes in ID patients compared to controls. Statistical follow-up approaches such as the case-control enrichment analyses that I will describe in chapter 3 can be a valuable method to give insights into the genetic aetiology of developmental disorders such as ID.

In chapter 4 I will describe a project in which two candidate dystroglycanopathy-associated genes, *B3GALNT2* and *GMPPB*, were identified using exome sequencing as part of the UK10K project. The aim of my work was to make zebrafish models of dystroglycanopathy by inhibiting the expression of each of these genes, and then to determine the extent to which the phenotype of these models recapitulated the phenotypes of the patients. I will demonstrate that zebrafish are an appropriate model for this purpose, and I will show that modelling candidate genes in zebrafish embryos is a functional follow-up approach that can help to determine whether a candidate gene is truly associated with a developmental disorder, and to give further insights into the pathology of that disorder.

The zebrafish project described in chapter 4 was carried out first (May 2011- February 2013), closely followed by the abnormal fetal development project described in chapter 2 (September 2011- November 2013) and then the project on the ID group of the UK10K project, described in chapter 3 (June 2013-August 2014). In this dissertation, I have described these projects in a non-chronological order, because the parts I played in each project flow more logically in the order in which I present them here. That is, for the abnormal fetal development project I was responsible for the majority of the analysis and interpretation of the exome sequencing data itself, for the ID project I was responsible for data analysis and also further statistical follow-up investigations, and for the zebrafish project I was responsible for functional follow-up of exome sequencing results using an animal model. All three projects serve to emphasise the importance of NGS for the diagnosis of rare developmental disorders, and for the identification of causal variants in novel genes.

2 Exome sequencing improves genetic diagnosis of structural fetal abnormalities

2.1 Introduction

2.1.1 The impact and causes of fetal structural abnormalities

The incidence of congenital abnormalities in the UK is approximately 2.2% (18). These are frequently first identified by ultrasound scan during the pregnancy. There is a wide range of potential outcomes for fetuses with abnormalities. Some abnormalities, such as isolated cleft lip, can be corrected in early childhood with a simple surgical procedure, and often has minimal long-term impact on the child (19). Others abnormalities, such as cerebral malformations, are associated with high morbidity and mortality (20).

Numerous genetic variants have been associated with fetal structural abnormalities. These include aneuploidies, copy number variants (CNVs), loss of function (LOF) single nucleotide variants (SNVs) and missense SNVs (21-23). Knowing the cause of a fetal structural abnormality can help clinicians to make an accurate prognosis regarding the pregnancy, and estimate recurrence risk for any future pregnancies. This helps the families to make informed decisions, including whether to terminate the pregnancy. Despite the importance of a diagnosis, currently only a minority of fetuses affected by a developmental disease receive a genetic diagnosis, to the frustration of families, clinicians and researchers alike (9).

2.1.2 Current techniques for prenatal genetic diagnosis

Sampling methods

Fetal DNA for genetic testing may be obtained invasively, by transabdominal or transcervical penetration of the uterus with a needle, in order to collect cells such as amniocytes or chorionic villus cells, from which fetal genomic DNA can be extracted. The major disadvantage of invasive sampling is that the risk of miscarriage increases by around 1% following a procedure (24). Also, sometimes a fetus and placenta may be mosaic for a particular mutation. That is, some of the cells carry the mutation and some do not. Therefore, another disadvantage is that if chorionic villus sampling is performed, and by chance only cells without the mutation are collected, the mutation will not be detected.

Alternatively, fragmented cell-free DNA (cfDNA) can be obtained non-invasively from maternal plasma; a proportion of this is fetal-derived (25). There are limitations to the application of this in prenatal diagnostics, as I will explain.

Karyotyping

One invaluable tool for the detection of chromosomal aberrations that cause fetal and congenital abnormalities is chromosome karyotyping, where whole chromosomes are stained and examined using a microscope. In classical cytogenetics, the stains (such as Giemsa stain) reveal patterns of light and dark bands that are unique to each chromosome. The technique was developed in the late 1960s, and it allowed researchers to distinguish between chromosomes of similar sizes for the first time (26). As karyotyping provides information on the number and gross appearance of chromosomes, it can be used to detect potentially pathogenic chromosomal aberrations including aneuploidy, deletions, duplications, inversions and translocations. Giemsa banding has a highest resolution of 3-10 Mb (27).

An alternative to classical cytogenetics is molecular cytogenetics, such as fluorescence *in situ* hybridisation (FISH). During this technique, fluorescent-tagged oligonucleotide probes complementary to a DNA sequence of interest are used to visualise whole chromosomes. It was first developed in the 1980s (28), and subsequent developments include chromosome ‘paints’ based on unique, chromosome-specific sequences which allow each chromosome to be visualised simultaneously in a different colour (29). Known as spectral karyotyping, this has some advantages over Giemsa banding in that it allows easy identification of the chromosomal origin of genetic material, and it has a higher resolution of 1-2 Mb (30). However, it is usually used in conjunction with other methods, as it has the major disadvantage of not being able to detect intrachromosomal aberrations.

FISH with locus-specific probes can identify known aberrations that cause fetal or congenital abnormalities. For example, 7q11.23 deletions in Williams syndrome, and dystrophin variants in Duchenne muscular dystrophy (31, 32). In another nice example of the clinical use of FISH, specific telomeric probes were used to identify an unbalanced subtelomeric translocation in a child with multiple congenital abnormalities, where classical cytogenetic analysis had indicated a normal karyotype (33). Generally, fetal chromosome karyotyping is offered to families when a significant fetal anomaly is identified by ultrasound, or when there is a high risk of such an anomaly. In these populations, karyotyping identifies a chromosomal anomaly in around 9% of cases (34).

Microarrays and quantitative fluorescent PCR

DNA microarrays include single nucleotide polymorphism (SNP) arrays and array comparative genomic hybridisation (aCGH). SNP arrays can be used for genotyping, identifying regions of absence of heterozygosity, performing genetic linkage analysis, and detecting unbalanced genomic rearrangements. aCGH can be used to detect CNVs that may be pathogenic, benign, or of unknown significance.

Microarrays have a higher resolution than G-band karyotyping. aCGH can detect deletions or duplications as small as 1 kb, depending on the platform used (35). A typical SNP array has a lower resolution of around 150-200 kb (36). For clinical diagnostic purposes, microarrays with a resolution in the range of 10-400 kb are considered to be the most cost-effective (37). An advantage of SNP arrays over aCGH is that they can be used to detect copy number neutral loss of heterozygosity, such as is caused by uniparental disomy. To utilise the advantages of both approaches, many modern platforms use both SNP probes and copy number probes on the same microarray.

One limitation of microarrays is that they are only able to detect unbalanced chromosomal rearrangements. Furthermore, they may not detect triploidy or low-level mosaicism (34, 38). Despite the limitations, microarrays have been the diagnostic test of choice for several years in children and adults with developmental delay (39). For fetuses with structural abnormalities, microarrays have a diagnostic yield of approximately 6-10% higher than chromosomal karyotyping (22, 34, 40).

Quantitative fluorescent polymerase chain reaction (QF-PCR) is an alternative method, during which amplification of repetitive loci is used to determine chromosomal copy number. QF-PCR is a cost-effective and robust method, which avoids the need to culture fetal cells, thus reducing turnaround time and eliminating the problem of introducing mutations during the culturing process (41). Because of these advantages, QF-PCR is now the clinical diagnostic test of choice for prenatal aneuploidy in the UK National Health Service (42).

Non-invasive prenatal testing

Between 3 and 50% of cfDNA in the plasma of a pregnant woman is fetal-derived (43-45). It consists of DNA fragments with a size range of 30-510 base pairs (bps), and a median of 162 bps (46). The cfDNA can be obtained non-invasively; therefore in recent years there has been huge interest in using it for prenatal genetic diagnosis. Non-invasive prenatal testing (NIPT) refers to assaying cfDNA to identify genetic variants in the fetus. This technique can be used to detect autosomal trisomies, sex chromosome aneuploidies, CNVs, fetal sex, rhesus status, and single gene disorders such as achondroplasia (34, 45, 47-49).

Regarding clinical practice, in the United States and China, use of NIPT to detect aneuploidies and fetal sex is already widespread (50, 51). Implementation for single-gene disorders is much slower because of lower demand and higher technical challenges. In the UK, NIPT is currently only being provided by the National Health Service for sex determination and some single-gene disorders. However, the RAPID study is investigating how to expand the implementation, and UK health professionals and parents generally view NIPT positively, therefore it is likely that provision will be expanded to other genomic disorders in the near future (52).

Two proof of concept studies published in 2012 showed that it is possible to sequence the whole genome of a fetus non-invasively using cfDNA, to a sufficient depth to be able to call inherited SNVs, using parental haplotypes to distinguish fetal from maternal variants (53, 54). However, the sensitivity and specificity of the SNV calling are as yet insufficient to consider using this approach in clinical practice.

For prenatal genetic diagnostics, it is very important to be able to identify *de novo* mutations, as they are often the cause of rare developmental phenotypes (11, 55-58). To detect *de novo* mutations non-invasively requires sequencing the cfDNA to a very high depth, because only a small proportion of fragments will carry the variant fetal allele. This is possible on a single-gene basis (49), but it is not currently possible genome-wide, at least not with any reasonable degree of sensitivity and especially specificity (54). Therefore, to identify potentially pathogenic SNVs and insertions or deletions (indels), on a large scale including those that occur *de novo*, in fetuses with structural abnormalities, next generation sequencing (NGS) on fetal DNA obtained through invasive methods remains, for now, the superior choice.

2.1.3 Next generation sequencing

NGS is a method of high-throughput DNA sequencing, which allows large amounts of genomic data to be generated quickly, and at a relatively low cost. The whole genome of an individual can be sequenced, or alternatively, particular genomic regions can be selected for sequencing, for example the exome, or diagnostic gene panels.

Exome sequencing is often favoured over whole genome sequencing, as it targets only coding regions, which represent 1-2% of the entire genome, but is said to contain up to 85% of the variants that cause known genetic disorders (59). Therefore exome sequencing is an efficient tool for gene discovery and genetic diagnostics in terms of cost, time and computational resources. The first report of exome sequencing as a method to discover the genetic cause of a Mendelian disease was made in 2010, with the identification of variants in *DHODH* as the cause of Miller syndrome (7). In the few short years since then, exome sequencing has proved to be a remarkably fruitful research tool, particularly for rare disease-associated gene discovery. At least one hundred genes that harbour variants causing Mendelian disease have been discovered, and this rate of progress shows no signs of abating as yet (8).

NGS is increasingly being used in the clinical setting, as a diagnostic test for patients with rare diseases. Often, exome sequencing is used. However, the most appropriate method depends upon the phenotype. For example, retinal dystrophy (RD) is a rare, inherited, degenerative cause of visual impairment and blindness. It is genetically heterogeneous, but a higher proportion of RD-associated genes have been identified, than for other phenotypes. Sequencing of 105 RD-associated genes therefore has a diagnostic yield of 55% (60). In contrast, exome sequencing of patients with rare, undiagnosed, developmental diseases typically has a diagnostic yield of around 25% (11, 61). Therefore, for phenotypes like RD, NGS using gene panels might be a more cost-efficient diagnostic method than exome sequencing.

Recently, as the cost of NGS has continued to fall, the prospect of using whole genome sequencing for rare disease-associated gene discovery and diagnostics has arisen. A recent study found that whole genome sequencing of patients with intellectual disability, for whom no likely cause of disease had been found by exome sequencing, had an impressive diagnostic yield of 42%, on top of what had been achieved by exome sequencing (62). This improvement was driven primarily by discovery of variants in coding regions that had been missed by the initial exome sequencing. Another recent study demonstrated that whole genome sequencing has more even coverage, and less bias in variant calling, than exome sequencing (63).

2.1.4 Variant prioritisation strategies

Interpretation of the tens of thousands of variants that are identified by NGS is challenging. A variant causing a rare, Mendelian disease must be rare in the general population. It is also likely to affect the structure or function of the protein encoded by the gene. Therefore, filtering the variants for rare, coding variants, along with various quality filters, is usually the first step in interpretation. The expected mode of inheritance of the disease is also taken into account. For example, if there is no family history of disease, variants with genotypes consistent with a *de novo,* recessive or X-linked (in the case of males) mode of inheritance will be prioritised. Of course, this requires that samples from parents are also available, which is not always the case. This basic filtering framework is the standard approach for both diagnostic and research applications (3, 7, 11), however it still often yields multiple candidate variants.

The next step depends on whether the application of the sequencing is clinical diagnostics, or research. For clinical diagnostics, matches between a gene that contains a variant in the patient, and genes that are known to be associated with the phenotype of that patient, are identified. For research, novel disease-associated genes are often identified by means of a functional link between a candidate gene and the phenotype of the patient. Some studies have attempted to partially systematise this inherently subjective approach using decision trees (11, 57). However, this approach is predicated on current knowledge of gene function, which for many genes is in its infancy. Thus, due to the subjectivity involved, there is a risk that the presence of *any* link between gene function and the phenotype could lead a researcher to ascribe pathogenicity to that variant. This approach is insufficiently stringent. For example, a recent paper looked at many genes in which variants are claimed to cause X-linked disability, and have found that several are in fact unlikely to be causative, because since the publication of the original studies, the patients' variants have been identified in control cohorts (64). It is imperative that a strict and consistent set of criteria for ascribing causality to a variant is developed and implemented across the rare disease genomics community to avoid such cases (12). To claim to have identified a novel disease-associated gene, recurrence of variants in multiple similar families over and above what might be expected by chance is usually also required.

There has been a lot of research in recent years into computational approaches for variant prioritisation. The main application of these is in novel disease-associated gene discovery rather than clinical diagnostics. Computational approaches have two obvious advantages over manual approaches. First, they are more objective and less biased, and second, they can prioritise much larger numbers of candidate variants than manual methods can.

The most basic methods are scores that indicate the probability that a variant is pathogenic based on various factors. For example, the PolyPhen and SIFT scores for missense variants are based on predicted degree of disruption to protein structure, and the evolutionary conservation of the amino-acid change. The GERP score is based on evolutionary conservation of a site, and the haploinsufficiency score is based on the probability that the gene is haploinsufficient (65-68). More advanced methods prioritise genes based on integrating different sources of information. Many such tools have been developed, and to name but one example Endeavour incorporates information on biological processes in which each candidate gene is involved (69).

2.1.5 Prenatal next generation sequencing: proof of concept

Because NGS can identify SNVs and indels throughout the genome, it has a much higher resolution than cytogenetic and array-based methods of variant discovery. Therefore, it is an obvious candidate method for prenatal diagnostics. Despite this, and despite the success of NGS in genetic diagnostics in rare disease postnatally, only a handful of studies have used it for prenatal gene discovery or diagnosis. The first two such studies, both published in 2012, used NGS to identify aneuploidy and chromosomal rearrangements. Dan *et al.* used very low-coverage whole-genome sequencing to detect aneuploidies and unbalanced chromosomal rearrangements in 13/62 fetuses (70), and Talkowski *et al.* used whole genome "jumping library" sequencing of amniocytes to identify an apparently balanced *de novo* translocation that disrupts *CHD7*, causing CHARGE syndrome in a single fetus (71).

The next two studies used exome sequencing at a depth sufficient to identify SNVs and indels, in a very small number of fetuses. Yang *et al.* performed exome sequencing on 250 patients with Mendelian disorders, four of which were fetuses from terminated pregnancies (11). In one of the fetuses, which had Cornelia de Lange syndrome, they found the cause of disease, which was a *de novo* splicing mutation in the known gene *NIPBL*. Finally, Filges *et al.* used exome sequencing to identify the cause of a recessive, lethal ciliopathy phenotype in one family (72). They sequenced the parents, their unaffected daughter, and post-mortem samples from two fetuses that were affected by the disease, and found compound heterozygous variants in *KIF14* in both affected fetuses.

2.1.6 Aims, context, and colleagues

Some parts of this project have been published (73, 74). The parts of these two publications that I have reproduced in this chapter were my work originally. This section briefly summarises the aspects of this study with which I was not directly involved, in order to put my own data into context.

The overall aims of this project were to use exome sequencing on a cohort of fetuses with structural abnormalities, and their parents, to estimate the diagnostic yield of this technique for this purpose, and to identify any issues that would need to be addressed prior to exome sequencing being implemented as a gene discovery or diagnostic tool for structural fetal abnormalities on a large scale.

A clinical team consisting of Dr Sarah Hillman, Dr. Dominic McMullan, Professor Eamonn Maher, and Professor Mark Kilby recruited a cohort of fetuses with structural abnormalities, and their parents, at the Fetal Medicine Centre Birmingham Women's Foundation Trust, UK. The fetal abnormalities were all first identified by ultrasound. The clinical team gathered further phenotypic data where available from post-mortem reports or paediatric follow up reports. Dr Sarah Hillman and Dr Dominic McMullan collected DNA samples from affected fetuses or neonates, and parental DNA. Prior to inclusion in this study the karyotypes were confirmed as normal, and low-resolution aCGH did not demonstrate any likely pathological CNVs.

The high-throughput sequencing team at the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK) did the exome sequencing itself. The Genome Analysis Production Informatics team at WTSI did the read mapping and variant calling. Dr Saeed Al Turki wrote Python scripts to calculate quality control metrics, and to identify and filter inherited variants, and he kindly allowed me to use them for this project. Dr Vijaya Parthiban developed the CoNVex program, and used it to identify CNVs from the exome data. Mr. Alejandro Sifrim developed the eXtasy program and ran it on these exome data, and Dr Damian Smedley developed PhenoDigm and ran it on these data.

The parts of this project that I was responsible for included assessing the quality of the exome data, analysing the data to identify rare coding variants consistent with the expected model of inheritance, designing a decision tree to use as a tool to interpret the variants, and (in close collaboration with the clinical team) interpreting the variants to decide which are likely causative. I carried out this work as described below, under the supervision of Dr Matthew Hurles.

2.2 Methods

2.2.1 Cohort

This cohort of 30 fetuses (that was collected, phenotyped and sampled by the clinical team at the University of Birmingham as described in section 2.1.6) is a subgroup (12%) of a larger cohort described previously (22). In this chapter, the participants are identified by their trio number prefaced by F for the fetus, M for the mother and P for the father. There are two exceptions to this, as the cohort includes two sets of related fetuses. F3 and F16 are monozygotic twins; therefore the parents of F16 are M3 and P3. F27 and F33 are siblings; therefore the parents of F33 are M27 and P27. F2 has an older sibling with a similar phenotype, who is not included in this study. The remaining fetuses are sporadic cases, and none of the parents had phenotypic abnormalities that were likely to be related to that of the fetuses. The trio numbers go up to 33, because there were originally 33 trios intended for sequencing, but exome sequencing failed due to insufficient DNA in trios 4, 24 and 30. The total cohort described here therefore consists of 26 trios and two quads (couple with two affected fetuses), which is a total of 30 affected fetuses.

The fetuses had a wide range of structural abnormalities (Figure 2-1). The three most commonly affected systems are the skeleton, the cardiovascular system and the nervous system. Abnormalities of skeletal morphology, such as agenesis of long bones, hemivertebrae, polydactyly, or talipes, were common in our cohort. Eighteen of the fetuses (60%) had at least one cardiovascular abnormality, such as ventricular septal defect, small heart, or defects of the valves or great arteries. Central nervous system defects included ventriculomegaly, and hypoplasticity of specific brain regions such as the cerebellum or the frontal lobe. Several of the mothers had abnormalities of the amniotic fluid such as anhydramnios or oligohydramnios, and five fetuses (17%) had generalised growth delay. Some fetuses (e.g. F1 and F10) had a very multisystemic phenotype, while others (e.g. F7 and F25) had a more specific phenotype, with a single affected system. Importantly, some of the fetuses underwent more extensive phenotyping (such as a post-mortem) than others. A detailed description of the phenotype of each fetus is recorded in the supplementary material of Carss *et al*. (73).

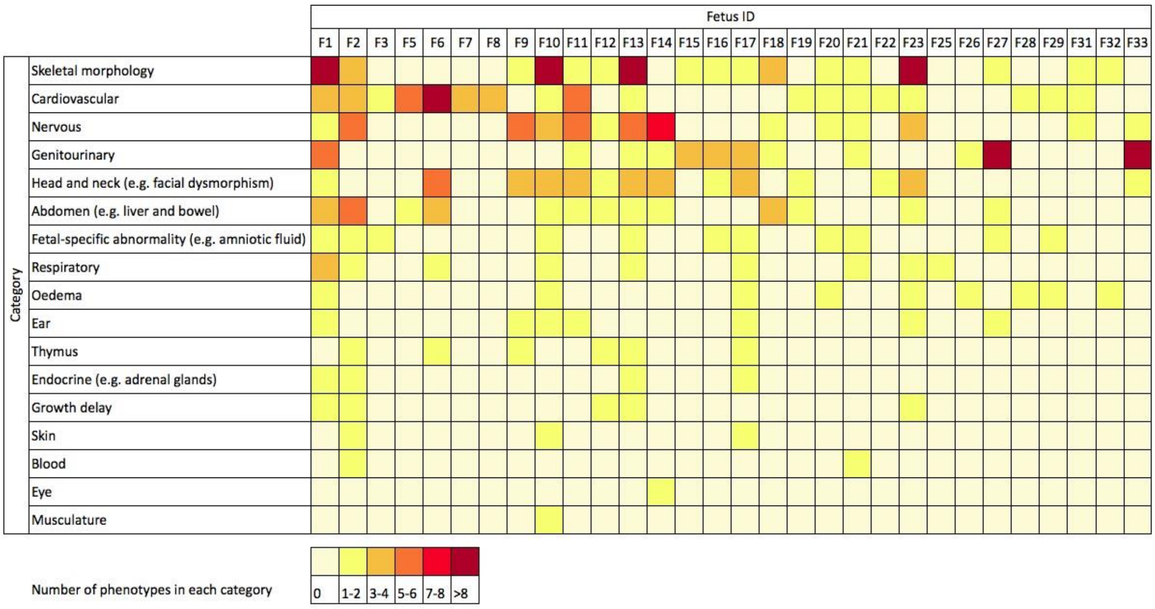

Figure 2 - 1: Matrix showing categories of phenotypes in the cohort of fetuses with structural abnormalities
For each fetus (F1-F33), the colour indicates the number of observed phenotypes that are in each category of phenotypes. For example, F1 has more than eight separate abnormalities of skeletal morphology. The categories are modified higher-order Human Phenotype Ontology (HPO) terms (75), and the data come from ultrasound scans, post-mortem reports or paediatric follow-up. This figure and legend have been published (73).

2.2.2 Exome sequencing

The DNA samples were sent to WTSI. Exome sequencing was performed using a SureSelect All Exon capture kit (50 Mb) version 3 (Agilent, Wokingham, UK), followed by paired-end sequencing (75 bp reads) on the HiSeq^{TM} platform (Illumina, Saffron Walden, UK). This work was done through an optimised pipeline run by the high-throughput sequencing team at WTSI. Reads were mapped to reference human genome GRCh37 (hs37d5). Variants were called using three different callers: SAMtools, GATK, and Dindel (76, 77). The Genome Analysis Production Informatics team at WTSI did this work.

2.2.3 VCF file merging, annotation, and quality control

For each of the samples, I merged the variant call format (VCF) files from the different variant callers using VCFtools (78). I added the following annotations to the VCF files:

gene name, variant consequence, PolyPhen score, and SIFT score using the Ensembl Variant Effect Predictor v2.2, and allele frequency information from 1000 Genomes Project (20101123 sequence release) (65, 66, 79, 80). I calculated quality control metrics using a Python script written by Dr Saeed Al Turki.

2.2.4 Identification of *de novo* SNVs and indels

To identify *de novo* mutations I used *De Novo* Gear pipeline version 0.6.2., which incorporates version 0.2 of *De Novo* Gear itself (41, 81). I used a two-tier strategy to filter the variants called by *De Novo* Gear. For genes not known to cause developmental disease (identified using the Developmental Disorder Gene2Phenotype (DDG2P) gene list available at https://decipher.sanger.ac.uk) I filtered out variants with minor allele frequency >0.01, in non-coding regions, depth <10x (in any member of the trio), in a tandem repeat or segmental duplication, I removed variants which occur in >10% of reads from either parent, and those where the calls in the VCF files were not consistent with a *de novo* mode of inheritance. Finally I visually inspected plots of the reads using the Integrative Genomics Viewer (IGV) and removed variants associated with reads that appeared to be incorrectly mapped (82). For genes in DDG2P I used a slightly less stringent filtering process to increase sensitivity. I removed variants with minor allele frequency >0.01, in non-coding regions, and those that appeared incorrectly mapped on IGV plots.

To calculate whether the final list of *de novo* mutations was enriched for functional mutations over what would be expected by chance, I calculated that the proportion of *de novo* mutations in exons expected to be functional by chance is 71.4% (83). I compared this to the proportion of *de novo* mutations that are functional in our cohort using a binomial test. To calculate the probability that a given number of functional *de novo* mutations will occur in the same gene in this cohort by chance, I calculated the number that are expected to occur using the known exome mutation rate, and the proportion of mutations that are expected to be functional, taking into account the length of the coding sequence of the gene of interest (83, 84). I compared this to the observed number of such mutations.

2.2.5 Identification of inherited recessive and X-linked SNVs and indels

I identified inherited SNVs and indels under different Mendelian models using Python scripts written by Dr Saeed Al Turki. This work was done twice. There was a preliminary round of analysis, then a final round of analysis, using improved filtering criteria (as described below).

For the preliminary round of analysis, I considered only variants that passed quality filters, were functional (predicted protein consequences were essential splice site, stop gained, complex indel, frameshift coding, non synonymous, stop lost), and had an allele frequency of <0.01 in the UK10K twins dataset (V4), the National Heart, Lung, and Blood Institute's Exome Sequencing Project (ESP, release ESP 6500_MAF_Jun_2012), and dbSNP. I also only considered variants in which the genotypes of the three members of the trio were consistent with inherited recessive (homozygous or compound heterozygous) or X-linked model of inheritance (in male fetuses), with unaffected parents.

For the final round of analysis, I made the following changes to the preliminary filtering protocol I have described. I no longer considered complex indels as candidates. This is because in between the preliminary and final rounds of analysis, the Ensembl variant effect predictor (VEP) was updated to version 68, which had improved methods to annotate the consequences of indels, and updated ontology for indels. Also, I considered only variants with an allele frequency of <0.01 in both the 1000 Genomes project, and an internal control cohort of 2172 individuals exome sequenced at the same laboratory, using the same pipelines and analysis methods. This is because using the internal cohort filter increased the specificity of the filtering, and not using the ESP and dbSNP databases may increase sensitivity, because these databases contain some disease-causing variants (85, 86).

2.2.6 Identification of CNVs

CoNVex detects copy number variation from exome data using comparative read depth. (ftp://ftp.sanger.ac.uk/pub/users/pv1/CoNVex/) It corrects for technical variation between samples and detects copy number variable segments using a heuristic error-weighted score and the Smith-Waterman algorithm. It detects deletions and duplications of targeted sequences from few hundred base pairs in size to a few megabases or more.

Dr Vijaya Parthiban ran CoNVex on this cohort. To identify candidate CNVs I filtered the CoNVex initial output. I considered only CNVs with CoNVex confidence score >=10, overlap within known common CNVs < 0.5, internal frequency of CNV in the dataset <0.05, overlaps at least one protein-coding gene, covered by >1 probe, and are not in an excessively noisy sample. I identified putative *de novo* and inherited X-linked CNVs in the fetuses, and inspected plots of regional $\log_2$ ratios in the family members and filtered out likely technical artifacts.

2.2.7 Sanger sequencing

I whole genome amplified ~50 ng genomic DNA from each sample using Illustra Genomiphi V3 ready-to-go kit (GE Healthcare Life Sciences, Buckinghamshire, UK) according to the manufacturer's instructions. I used this as a template to amplify a fragment containing each the variant of interest in the relevant trios using REDTaq® DNA Polymerase (Sigma-Aldrich, Dorset, UK) and capillary sequenced using BigDye v31 kit and ABI 3730 sequencer according to the manufacturers' instructions. Primers that were used to validate variants are listed in Appendix 1.

2.2.8 Interpretation of variants

To interpret the variants, I first annotated each candidate gene with functional information (where available) from the databases listed below.

- OMIM (http://www.omim.org/)
- DDG2P (http://decipher.sanger.ac.uk/ddd/ddd_genes)
- BioGPS (biogps.org)
- NHGRI GWAS catalog (http://www.genome.gov/gwastudies/)
- IKMC (http://www.knockoutmouse.org/)
- ZFIN (http://zfin.org/)
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed)

I next developed and used a decision tree to classify each variant as being highly likely to be causal, possibly causal but requires further genetic or functional confirmatory studies, or unknown (Figure 2-2). This work was done in close collaboration with the clinical team at the University of Birmingham. Mr. Alejandro Sifrim developed the eXtasy program and ran it on these exome data, and Dr Damian Smedley developed PhenoDigm and ran it on these data. To calculate the 95% confidence interval limits for my estimate of diagnostic yield, I used a binomial test.

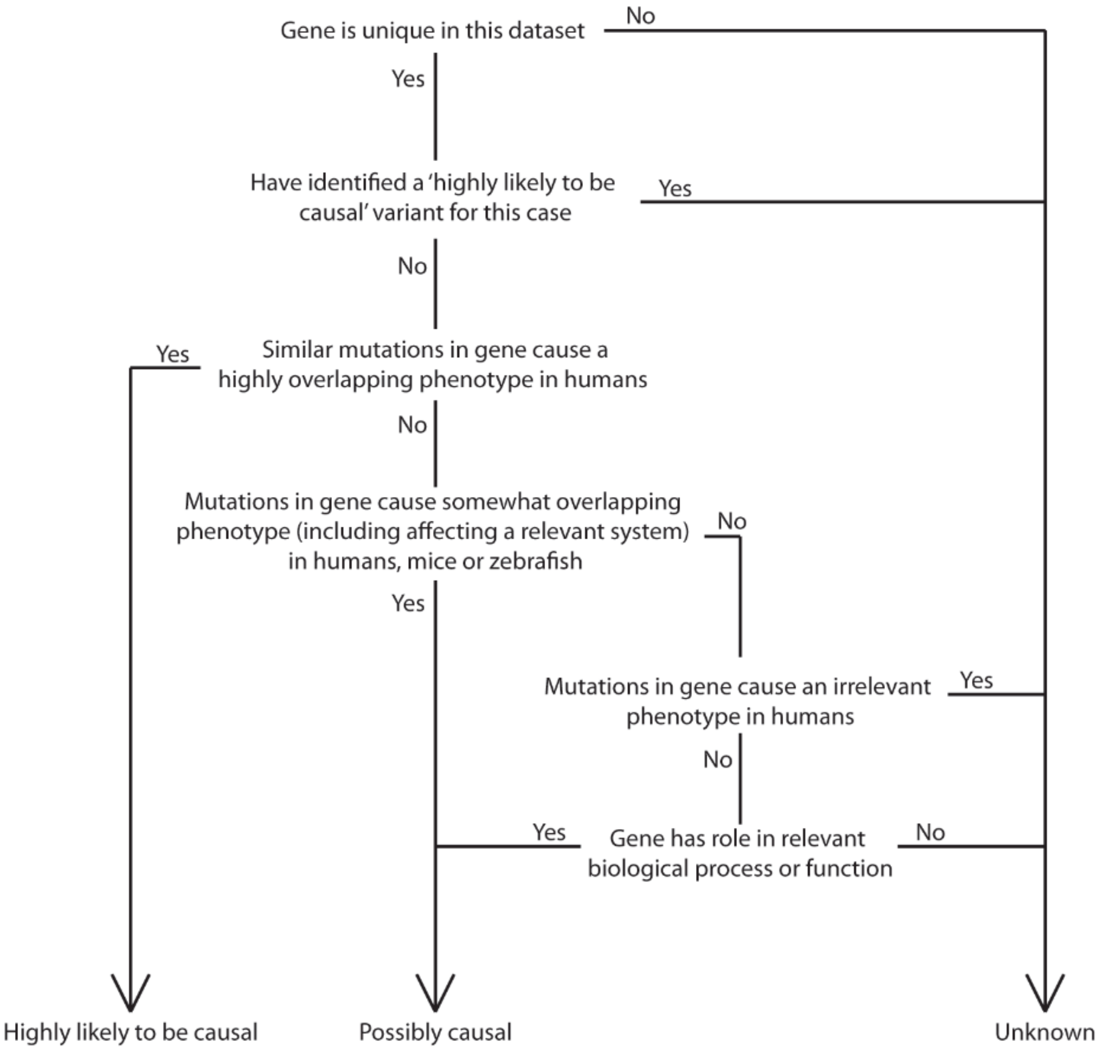
Gene is unique in this dataset
No
Yes
Have identified a 'highly likely to be causal' variant for this case
Yes
No
Similar mutations in gene cause a highly overlapping phenotype in humans
Yes
No
Mutations in gene cause somewhat overlapping phenotype (including affecting a relevant system) in humans, mice or zebrafish
No
Yes
Mutations in gene cause an irrelevant phenotype in humans
Yes
No
Gene has role in relevant biological process or function
Yes
No
Highly likely to be causal
Possibly causal
Unknown

Figure 2 - 2: Decision tree for classifying candidate genes into three categories.
Data were used where available from the following sources: Online Mendelian Inheritance in Man (OMIM), DDG2P, Biology Gene Portal System (BioGPS), National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalogue, International knockout mouse consortium (IKMC) database, zebrafish information network (ZFIN) database and PubMed. This figure and legend have been published (73).

2.3 Results

2.3.1 The exome sequencing data are of high quality

Exome sequencing in 30 fetuses and neonates with a diverse range of structural abnormalities diagnosed at prenatal ultrasound, along with their parents, was performed (a total of 86 individuals). The mean depth of coverage of the targeted coding regions was 103X. This coverage is much higher than the minimum 30X estimated to be required for accurate detection of heterozygous variants (87). A mean of only 7.3% of bases in the targeted coding regions had less than 10X coverage, and a mean of only 1% had less than 1X coverage (Figure 2-3 and Table 2-1).

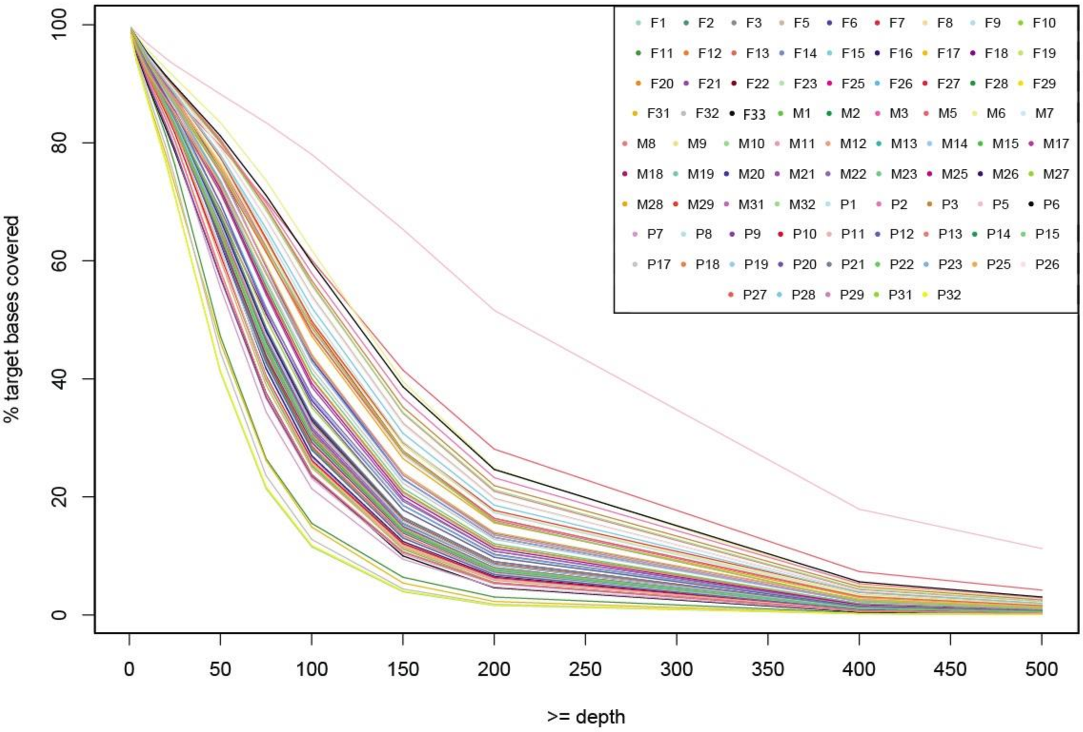

Figure 2 - 3: Target coverage of exome sequencing reads by sample.
P5 has higher coverage, as it was not sequenced as part of a pool. This figure and legend have been published (73).

ID	N mapped HMQ reads	% Q20 bases	Mean coverage	>=1x (%)	>=10x (%)	N coding variants
F1	71288090	95.71	106.274	99.25	94.11	21826
F2	71507176	95.75	106.564	99.03	93.53	21667
F3	76307901	95.68	114.432	98.95	93.11	21742
F5	92699881	95.66	137.72	99.42	95.02	21954
F6	75797156	95.83	113.295	99.16	94.3	21940
F7	84423053	95.6	125.512	99.18	94.36	21552
F8	84367449	95.64	125.866	99.37	94.78	21687
F9	83754651	95.7	125.248	99.25	94.49	21742
F10	53387862	95.8	79.831	98.78	92.15	21440
F11	40775602	95.85	61.05	98.62	89.9	20857
F12	53976303	95.75	80.52	98.75	91.81	21367
F13	57086795	95.82	85.211	98.95	92.23	21237
F14	55239595	95.76	82.435	98.98	92.82	21663
F15	58512496	95.76	87.287	98.78	92.05	21155
F16	55517406	95.68	83.102	99.2	93.09	21956
F17	56395887	95.77	84.406	98.82	92.42	21640
F18	66147741	96.59	98.053	98.62	91.17	20964
F19	58908353	95.53	87.821	98.92	92.07	21779
F20	70831428	96.63	105.403	98.95	92.37	21281
F21	68895929	96.67	102.558	99.07	92.73	21127
F22	56904719	95.5	84.907	99.06	92.78	21498
F23	58600063	95.45	87.365	98.82	91.95	21353
F25	59597856	95.49	88.807	98.95	92.04	21513
F26	66648868	96.6	99.192	98.84	92.11	20982
F27	59205640	95.53	88.366	98.84	92.44	21535
F28	62500308	95.43	93.196	98.85	92.54	21525
F29	76111740	96.6	113.526	98.93	92.97	21219
F31	38825777	96.56	57.866	98.17	87.64	20468
F32	37322925	96.44	55.537	98.46	88.61	21046
F33	49286255	96.58	73.419	98.64	89.88	21075
M1	52645663	95.44	78.481	98.84	92.15	21498
M2	59986920	95.54	89.333	98.81	92.15	21499
M3	82707758	96.16	123.515	98.99	93.98	21784
M5	110411666	95.87	165.606	98.73	93.19	21456
M6	104330917	95.66	155.316	99.36	95.82	22028
M7	82706691	96.1	123.255	99.16	94.58	21622
M8	95419993	96.03	142.143	99.17	94.99	21817
M9	85590849	96.18	127.864	98.94	93.72	21612
M10	95663391	96.13	142.556	99.16	94.78	21956
M11	50195030	96.2	75.003	98.48	90.88	20901
M12	52632594	96.17	78.669	98.67	91.75	21451
M13	60213492	96.13	90.143	98.51	91.41	21258
M14	58095399	96.11	86.586	98.89	92.49	21152
M15	56736873	96.19	84.692	98.85	92.43	20934
M17	60229898	96.09	89.792	98.86	92.76	21449
M18	57601439	96.18	85.68	98.74	91.84	20930
M19	58492263	96.15	87.242	98.86	92.63	22220
M20	62693258	95.86	93.795	98.72	92.12	21422
M21	60860845	95.9	90.91	98.67	91.94	21212
M22	67534892	95.84	100.657	98.64	91.98	21408
M23	72503603	95.8	107.912	99.05	93.61	21670
M25	69963332	95.77	104.385	98.97	93.42	21374
M26	62052636	95.85	92.442	98.74	92.31	21378

M27	65123188	95.9	97.14	98.72	92.33	21177
M28	81636876	97.13	121.798	99.1	93.95	21812
M29	86596684	97.08	130.433	99.14	94.3	21597
M31	81006172	96.57	120.641	99.34	93.57	21736
M32	35173157	96.5	52.393	98.15	87.36	20791
P1	85699745	96.98	128.031	99.32	94.54	21561
P2	100554052	97.12	150.412	99.39	94.93	21516
P3	97431748	97.2	145.569	99.42	95.07	21968
P5	174856038	96.96	260.148	99.58	96.84	21617
P6	103897082	95.79	154.839	99.37	95.17	21319
P7	48099786	97.81	71.891	98.73	90.69	21121
P8	55948619	97.83	83.722	98.84	91.48	21189
P9	50521398	97.75	75.181	98.73	90.39	21042
P10	54187949	97.78	80.854	98.94	91.66	21339
P11	53758221	97.73	80.136	99.1	92.28	21611
P12	56321179	97.73	83.975	99.05	92.25	21454
P13	51049757	97.78	76.128	98.8	91.15	21229
P14	58646676	97.8	87.73	98.98	92.27	21445
P15	59527162	97.43	88.824	99.01	92.44	21636
P17	73688831	97.47	110.281	99.18	93.41	21268
P18	61532376	97.39	91.506	98.95	91.71	21431
P19	61501500	97.39	91.594	99.11	92.53	21757
P20	65921431	97.39	98.197	99.19	93.29	21340
P21	62992323	97.4	94.03	99.04	92.52	21352
P22	58820342	97.44	87.688	98.96	92.12	21274
P23	75143669	96.28	111.818	99.22	93.66	21781
P25	76510093	96.69	114.498	98.76	92.09	21325
P26	92137474	96.27	137.371	99.36	94.84	21831
P27	82888871	96.22	123.255	99.22	93.7	21487
P28	89608716	96.23	133.306	99.36	94.64	21526
P29	76685316	96.29	114.252	99.32	94.22	21583
P31	81272187	96.53	120.872	99.5	93.76	21304
P32	35398578	96.59	52.7	98.31	87.25	20983

Table 2 - 1: Exome sequencing coverage and quality control metrics.
Numbers apply to target coding regions only. N = number; Q20 = Number of bases with a phred-like calibrated quality score of 20 or above; HMQ = high mapping quality (>Q30). This figure and legend have been published (73).

The mean number of SNVs detected per sample was 73970, of which 10329 were functional, 10623 were silent, 134 were LOF, and 94.5% were common (≥1% population frequency) (Figure 2-4). Of the LOF variants, only 86.6% were common. The mean transition/transversion ratio of SNVs was 3.014, which is close to the expected value (88). The mean number of indels per sample is 8722, of which 84% are common (Figure 2-5). The mean number of coding indels per sample is 449, of which 85.7% are common. The mean in-frame/frameshift ratio of coding indels is 1.47, because there is a bias towards less damaging in-frame indels. This is close to the expected value (89).

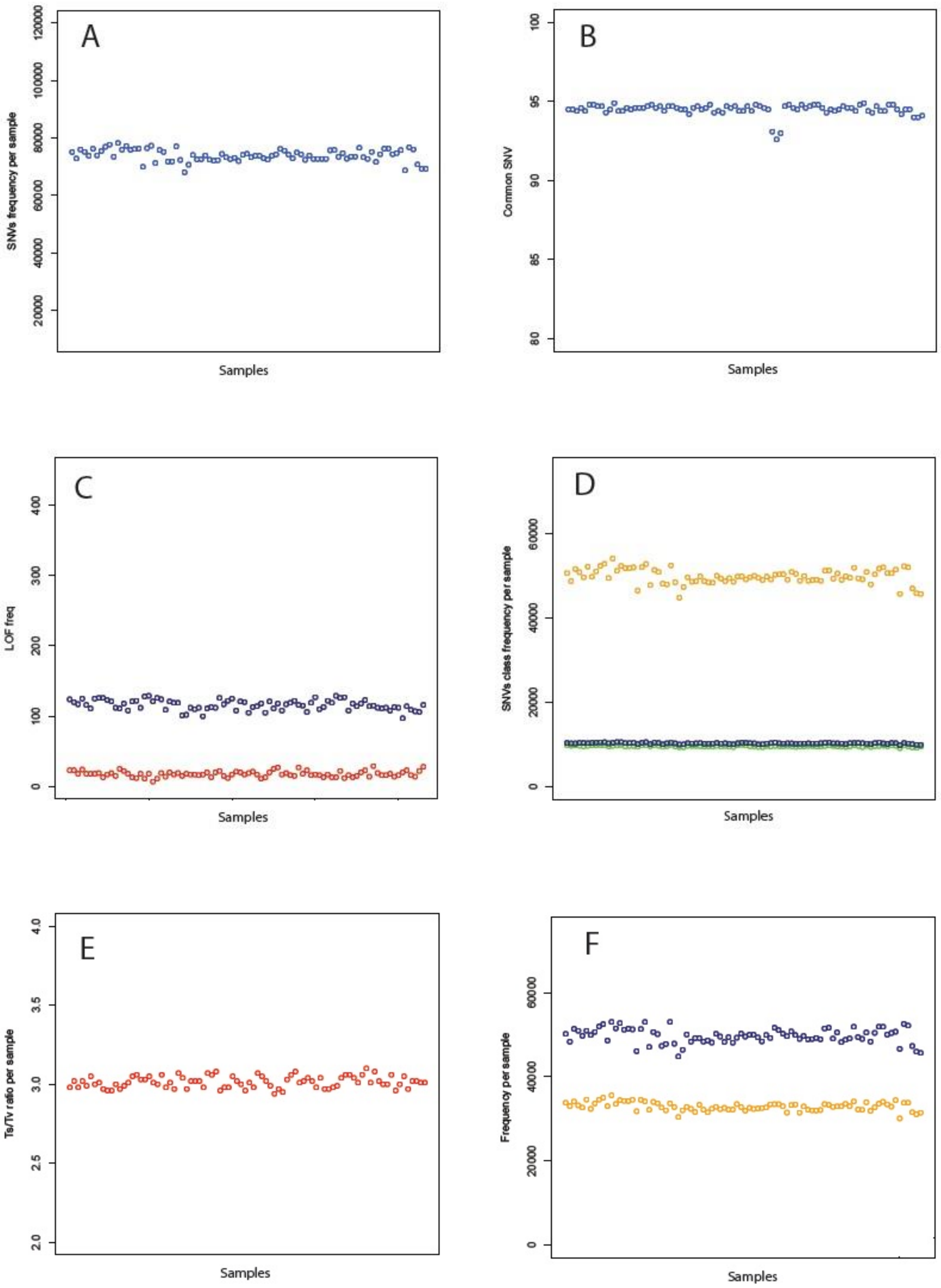

Figure 2 - 4: Quality control metrics for single nucleotide variants.
(A) Number of high-quality SNVs per sample. **(B)** Percent of SNVs that are common (≥1% population frequency) per sample. The cluster of three samples with a lower percentage of common SNVs represents F19, M19 and P19. These individuals are of Indian ancestry, whereas most of the cohort is of European ancestry. **(C)** Number of LOF SNVs per sample. Common (≥1%) are shown in blue and rare (<1%) are shown in red. **(D)** Number of SNVs per sample that are functional (green), silent (blue) and other (yellow). **(E)** Transition/transversion ratio per sample. **(F)** Number of SNVs per sample that are heterozygous (blue), and homozygous (yellow). This figure and legend have been published (73).

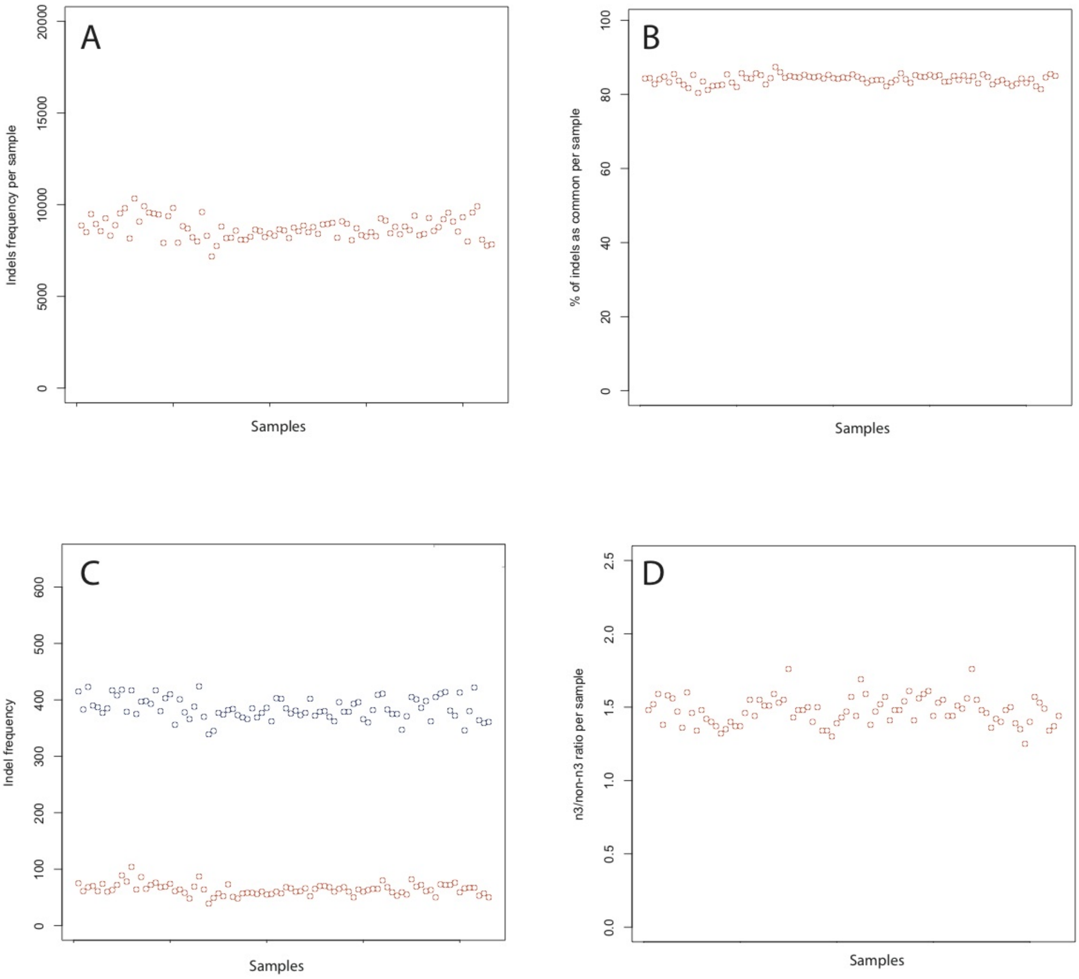

Figure 2 - 5: Quality control metrics for indels.
(A) Number of high-quality indels per sample. **(B)** Percent of indels that are common (≥1% population frequency) per sample. **(C)** Number of coding indels per sample. Common (≥1%) are shown in blue and rare (<1%) are shown in red. **(D)** Ratio of coding indels with length that is a multiple of three against coding indels with length that is not a multiple of three, per sample. This figure and legend have been published (73).

No parental phenotypic abnormalities were reported that might be related to the fetal abnormalities, suggesting dominant inheritance is unlikely. I therefore identified rare, coding variants under dominant *de novo*, recessive and X-linked (for male fetuses) modes of inheritance. No parental consanguinity was reported. Next, through systematic manual curation of the existing literature and databases, I classified the variants into one of three categories: highly likely to be causal, possibly causal, or unknown. For the three non-sporadic cases (the siblings F27 and F33, and F2, who has a similarly affected sibling not included in this study), all of which are female, I consider a recessive mode of inheritance most likely. I nevertheless investigated all the variant classes described above.

2.3.2 There is a mean of 1.13 validated *de novo* SNVs or indels per fetus

I identified potential *de novo* SNVs and indels with high sensitivity, and inevitably low specificity, yielding a list of 77 candidate *de novo* coding or splicing mutations (mean=2.6 per fetus, range = 0-5). I attempted to validate all of these by capillary sequencing of whole genome amplified genomic DNA, irrespective of their predicted functional consequence. I validated 34 as being truly *de novo* (Table 2-2). This is a mean of 1.13 per fetal exome (range 0-4), which is within the expected range from the known germline mutation rate, and NGS of other disease cohorts (56, 57, 84, 90). These mutations include identical *PPFIBP2* mutations in the monozygotic twins F3 and F16, with the result that there are 33 independent *de novo* mutations.

ID	CHR	POS	REF	ALT	Gene	CQ	N REF	N ALT	P
F2	16	9857047	G	A	*GRIN2A*	NS	29	24	0.29
F3	11	7618837	G	C	*PPFIBP2*	NS	28	16	0.048
F6	11	33677654	C	T	*C11orf41*	STOP	43	46	0.66
F6	12	56567575	G	A	*SMARCC2*	STOP	122	102	0.1
F6	17	29562669	G	A	*NF1*	NS	146	133	0.24
F6	20	39813788	G	A	*ZHX3*	S	9	4	0.13
F7	2	210694087	G	A	*UNC80*	NS	138	136	0.48
F7	20	44190748	C	T	*WFDC8*	SPLICE	28	30	0.65
F8	1	160811672	G	T	*CD244*	NS	33	38	0.76
F9	2	205829965	G	C	*PARD3B*	NS	79	25	5.3 x 10^{-8}
F10	8	20069263	G	T	*ATP6V1B2*	NS	26	20	0.23
F10	9	91994007	C	T	*SEMA4D*	NS	10	7	0.31
F14	1	28099859	C	T	*STX12*	NS	8	12	0.87
F14	4	44450177	C	T	*KCTD8*	NS	14	13	0.5
F15	10	128830000	G	A	*DOCK1*	NS	147	158	0.75
F16	11	7618837	G	C	*PPFIBP2*	NS	18	19	0.63
F18	3	58639419	G	A	*FAM3D*	NS	65	44	0.027
F18	12	123444538	G	A	*ABCB9*	NS	7	8	0.7
F19	2	205983695	G	A	*PARD3B*	NS	67	56	0.18
F19	3	132230069	T	C	*DNAJC13*	S	45	37	0.22
F19	17	5461819	G	C	*NLRP1*	NS	30	31	0.6
F20	12	48369853	C	A	*COL2A1*	NS	22	30	0.89
F22	10	71175853	G	A	*TACR2*	NS	11	16	0.88
F23	4	1806099	A	G	*FGFR3*	NS	57	42	0.08
F25	3	47727627	G	A	*SMARCC1*	STOP	17	15	0.43
F25	10	118359676	C	T	*PNLIPRP1*	NS	77	57	0.05
F26	1	202722193	C	A	*KDM5B*	NS	45	24	0.0077
F26	8	74334894	T	G	*STAU2*	NS	48	37	0.14
F27	2	106687405	A	G	*C2orf40*	NS	20	14	0.2
F27	11	15260600	G	A	*INSC*	NS	10	12	0.74
F28	19	55748185	T	C	*PPP6R1*	NS	27	29	0.66
F31	12	50047598	G	C	*FMNL3*	NS	38	24	0.049
F33	10	102249809	C	A	*SEC31B*	NS	21	5	0.0012
F33	X	13645272	G	A	*EGFL6*	S	111	92	0.1

Table 2 - 2: Validated *de novo* SNVs in fetuses with structural abnormalities.
ID = ID of fetus; CHR = chromosome; POS = position; REF = sequence of reference allele; ALT = sequence of alternate allele; CQ = consequence of mutation; NS = non-synonymous coding variant; S = synonymous coding variant STOP= stop codon gained; SPLICE = essential splice site variant; N REF = number of sequencing reads that support the reference allele; N ALT = number of sequencing reads that support the alternate allele; P = p value from binomial test to test whether the proportion of sequencing reads that support the alternate allele is significantly less than 0.5 (Bonferroni-corrected threshold of significance = 0.00147). This table and legend have been published (73).

The expected percentage of *de novo* mutations in coding or splicing sequence that are synonymous is 29% (83), however, I observed that only three (9%) of the 33 validated independent *de novo* mutations were synonymous, with 26 being non-synonymous, three nonsense and one in a splice site. Thus the proportion of validated *de novo* mutations that are predicted to have a functional consequence of the encoded protein is significantly enriched over what would be expected by chance (p=0.007), suggesting that an appreciable subset of these functional mutations is likely to be pathogenic. For two of the *de novo* mutations, the proportion of reads that support the alternative allele was significantly less than the expected 50% for a non-mosaic, heterozygous mutation. This provides suggestive evidence that these mutations are mosaic. These mutations were c.313G>C (p.105E>Q) in *PARD3B* (ENST00000349953) in F9, and c.2921G>T (p. 974C>F) in *SEC31B* (MIM 610258, ENST00000370345) in F33 (Table 2-2).

2.3.3 There are three candidate *de novo* or X-linked copy number variants

CNVs from the exome data were denoted using the CoNVex program. I identified three rare, high-quality CNVs (one deletion and two duplications) under *de novo*, inherited recessive, or X linked models (Table 2-3 and Figure 2-6).

ID	CHR	Start position	End position	Size (kb)	CNV type	Inheritance model	Gene
F14	X	13770686	13791294	20.6	DEL	*de novo*	*GPM6B;* *OFD1*

F19	X	48155306	48270940	115.6	DUP	Inherited X linked	*SSX3;*
							SSX4;
							SSX4B

F3	X	103267111	103301913	34.8	DUP	Inherited X linked	*H2BFM;* *H2BFWT*

Table 2 - 3: Candidate CNVs in fetuses with structural abnormalities.
None of the genes in these CNVs have additional variants likely to cause disease. None of these CNVs have any overlap with common CNVs.

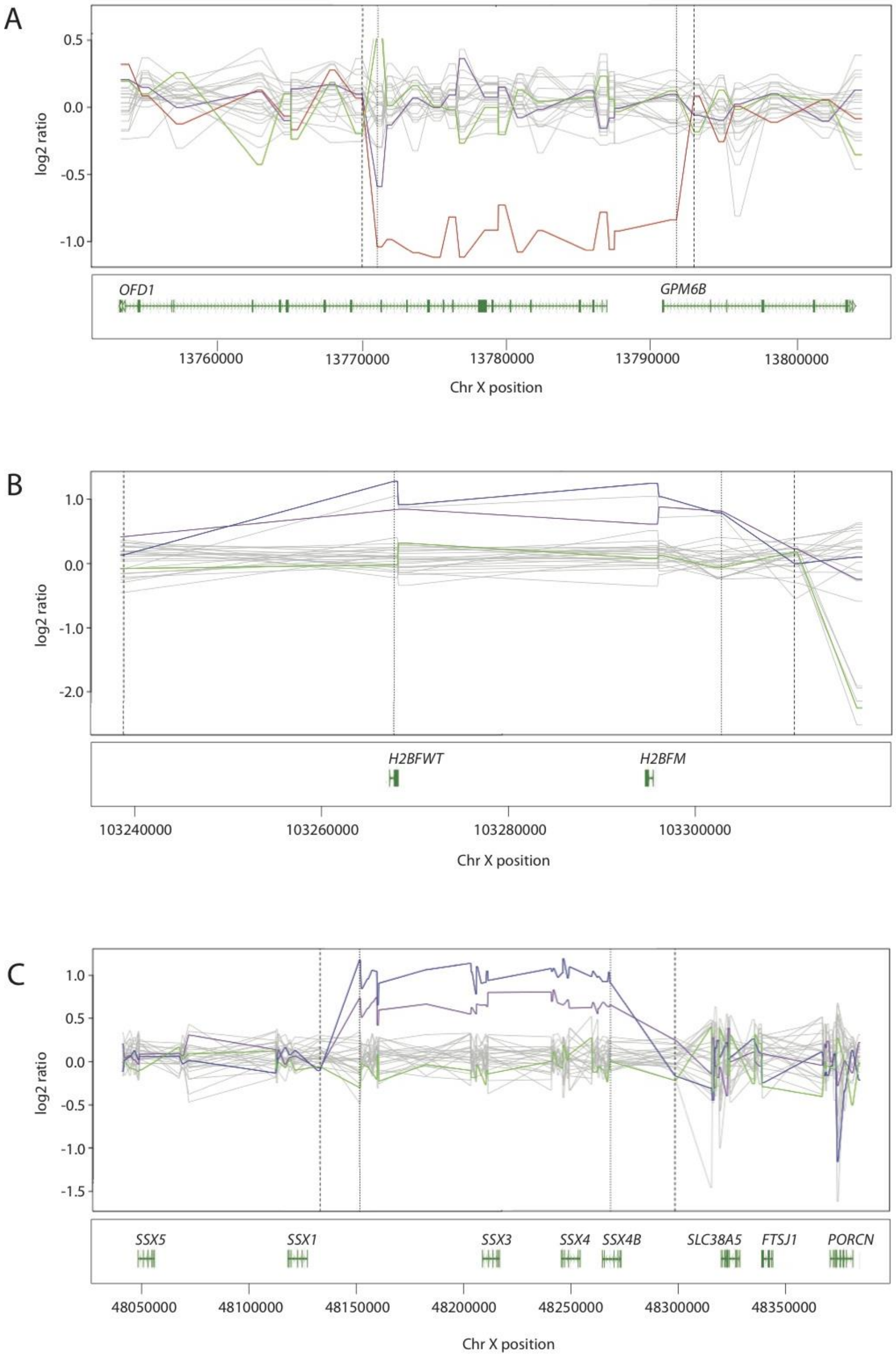

Figure 2 - 6: Log$_2$ ratios of candidate CNVs in fetuses with structural abnormalities.
(A) F14; **(B)** F19; **(C)** F3. In each plot the x-axis indicates the genomic coordinates. The top panel indicates the normalised log$_2$ ratio of the exome read depth, compared to a group of controls. The red line shows the log$_2$ ratio of the fetus, where the variant is a deletion, and the blue line shows the log$_2$ ratio of the fetus where the variant is a duplication. The purple line shows the log$_2$ ratio of the mother, and the green line shows the log$_2$ ratio of the father. The grey lines show the log$_2$ ratio of control samples. The vertical small dashed lines show the minimum deleted/duplicated region and the vertical wide dashed lines show the maximum deleted/duplicated region. The bottom panel shows the protein-coding genes present in each region. This figure and legend have been published (73).

2.3.4 There is a mean of 13 candidate genes with inherited recessive or X-linked variants per fetus, in the preliminary round of analysis

Identification and interpretation of inherited recessive or X-linked SNVs and indels was done twice in this project. There are three differences between these preliminary and final rounds of analyses. In the preliminary round, only samples F1-F30 were included, because samples F31-F33 were sequenced later, in a separate batch. Second, I used a slightly different, more sensitive and specific filtering protocol for the final round. Finally, for variant interpretation, in the final round I was able to take into account data from computational gene prioritisation methods, as I will describe.

For the preliminary round of analysis, I identified potentially relevant inherited recessive and X-linked variants (SNVs and indels) by filtering for rare (minor allele frequency less than 1%), functional hemizygous, homozygous or compound heterozygous variants. This identified a mean of 13 candidate genes per fetus (range of 6-21) with a cumulative total of 256 candidate genes across the 27 fetuses, containing 505 rare functional variants. Of these variants, 450 are missense, 40 are frameshift indels, 9 are in-frame indels and 6 are nonsense (Appendix 2). Of the candidate genes, 47 were observed in more than one individual in this cohort (not including the twins F3 and F16).

I next used my decision tree to categorise each variant in each of the three categories (*de novo* SNVs and indels, CNVs, and inherited SNVs and indels) as being highly likely to be causal, possibly causal, or unknown. This work was done in close collaboration with the clinical team at the University of Birmingham. In the following sections I describe the variants I categorised as highly likely to be causal or possibly causal in each category, and explain my rationale for these categorisations.

2.3.5 *De novo* SNVs in *FGFR3* and *COL2A1* are highly likely to be causal

Two of the *de novo* SNVs are highly likely to be pathogenic, and two are possibly causal. One *de novo* mutation that is highly likely to be causal was found in F23, a male fetus with features consistent with thanatophoric dysplasia, including a large head, disproportionately short limbs, and a narrow, bell-shaped chest. I found the missense mutation c.1118A>G (p.373Y>C) in fibroblast growth factor receptor 3 (*FGFR3*, MIM 134934 (http://www.omim.org/), ENST00000440486) (Figure 2-7). FGFR3 is a well-characterised negative regulator of bone growth, missense mutations in which are known to cause a wide range of skeletal dysplasias, most commonly achondroplasia. There is a very tight correlation between specific *FGFR3* mutations, and the phenotype, for a review see (91). The mutation p.373Y>C is known to cause thanatophoric dysplasia (23), giving high confidence that c.1118A>G in *FGFR3* is the causative mutation in F23.

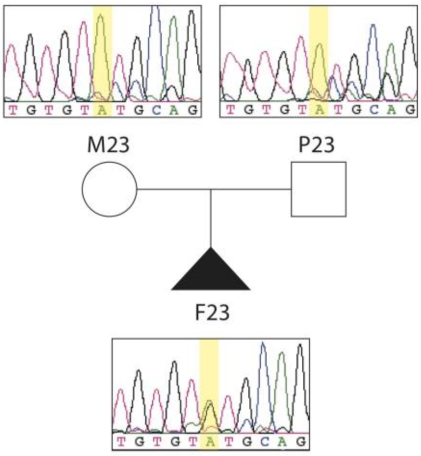

Figure 2 - 7: Pedigree of trio 23, showing Sanger sequencing of *de novo* mutation in *FGFR3*.

In F20, a male fetus with increased nuchal translucency (>3.5mm), tricuspid regurgitation, and an extended posture and bilateral talipes equinovarus anomaly I found the highly likely to be causal missense mutation c.3490G>T (p.1164G>C) in *COL2A1* (MIM 120140, ENST00000380518) (Figure 2-8). Mutations in this gene, which encodes COL2A1, a component of type II collagen, can cause type II collagenopathies. This term covers a wide spectrum of phenotypes, from the lethal achondrogenesis type II (MIM 200610) which typically involves very severe dwarfism with a short chest and can involve heart defects and structural defects of the lower limb (92, 93), to much milder phenotypes such as spondyloperipheral dysplasia (MIM 271700), which includes short stature and other skeletal defects such as talipes and other lower limb abnormalities (94). Importantly, p.1164G>C is a glycine to non-serine in the triple helical domain of COL2A1, which is predicted to be a particularly damaging class of substitution (95), although p.1164G>C has not previously been reported.

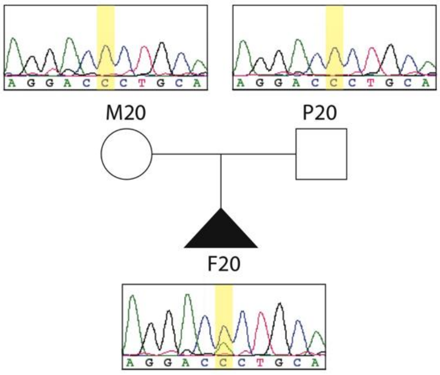

Figure 2 - 8: Pedigree of trio 20, showing Sanger sequencing of *de novo* mutation in *COL2A1*.

2.3.6 *De novo* SNVs in *NF1* and *SMARCC2* are possibly causal

F6 is a female fetus with levocardia with abdominal situs inversus, malposed great arteries, and multiple ventricular septal defects. Some of these features are consistent with Ivemark's syndrome (MIM 208530), the molecular basis of which is unknown. In this fetus I found three possibly pathogenic variants, two of which are *de novo*. I found the *de novo* mutation c.2747G>A (p.916R>Q) in *NF1* (MIM 613113, ENST00000456735). Variants in this gene, which encodes neurofibromin 1, most commonly cause neurofibromatosis, but in a subset of patients variants are associated with Neurofibromatosis-Noonan syndrome (MIM 601321), one feature of which can be cardiac defects including atrial septal defect (96). Mutation of this particular amino acid has been previously proposed to be pathogenic (97). Additionally, zebrafish knockdowns for either orthologue of *NF1* (*nf1a* or *nf1b*) have cardiovascular defects including valvular insufficiency (98).

In F6 I also found a nonsense mutation c.1555C>T (p.519R>*) in *SMARCC2* (MIM 601734, ENST00000267064). This encodes the SWI/SNF-related chromatin regulator SMARCC2 that, while not known to be associated with human developmental disease, does have a role in development (specifically differentiation of embryonic stem cells) (99). Heterozygous LOF variants within several genes that encode components of the same protein complex or family (such as *SMARCAL1*) can cause developmental disorders (58, 100). Similarly, I found a *de novo* nonsense mutation c.1297C>T (p.433R>*) in *SMARCC1* (MIM 601732, ENST00000254480) that I initially classified as possibly causal in F25. However, follow up of this case showed that the fetal phenotype (hydrothorax with mediastinal shift) resolved postnatally. Therefore this mutation, despite appearing possibly clinically relevant, is unlikely to be significantly pathogenic.

I looked for inherited, rare, coding, ‘second hit’ variants in genes in which I found *de novo* mutations and found only one: a heterozygous, maternally inherited, missense variant in *SEMA4D* in F10.

De novo mutations in genes known to be involved in developmental disease were not necessarily classified as possibly causal, where the phenotype of the fetus did not overlap sufficiently with previously reported phenotypes. For example, the *de novo* missense mutation c.4354C>T (p.1452R>C) in *GRIN2A* (MIM 138253, ENST00000461292) was found in F2, a female with atrioventricular septal defect (AVSD), hepatic dysfunction, polydactyly, panhypopituitarism and brain injury. *GRIN2A* mutations can cause seizures and intellectual disability, and are highly unlikely to be the cause of the multiple structural malformations seen in F2 (101). Supporting this assertion is the fact that this individual had an older sibling with a similar phenotype, making *de novo* mutations an unlikely cause of disease.

2.3.7 Two unrelated fetuses with no clear clinical overlap have *de novo* SNVs in *PARD3B*

Two of the unrelated fetuses had *de novo* missense mutations in *PARD3B*. F9, a male fetus with a complex brain malformation and unilateral talipes equinovarus had the *PARD3B* mutation c.313G>C (p.105E>Q). F19, a male with an atrial septal defect, oesophageal atresia and a unilateral facial cleft had the mutation c.731G>A (p.244R>Q). The likelihood of two functional *de novo* mutations in a gene of the size of *PARD3B* occurring by chance in unrelated probands in a cohort of this size is small (p = 3.1×10^{-6}), but does not quite reach the Bonferroni-corrected significance threshold for testing of all genes of p = 2.5×10^{-6}. *De novo PARD3B* mutations have not been reported in other larger sequencing studies suggesting that *PARD3B* does not have an unusually high mutation rate (57, 84). *PARD3B* encodes partitioning defective 3 homolog B (Par3b), which is involved in cell polarisation (102). It has a paralogue, *PARD3,* which has a role in various developmental processes including neurogenesis (103). Homozygous mouse knockouts for *Par3* are embryonic lethal and have growth retardation, heart and brain defects and short tails (104), and zebrafish *pard3* knockdowns have hydrocephalus (103). The overlap between phenotypes resulting from knockdown of *PARD3* and the phenotypes in F9 and F19 is interesting, however I judged that the current knowledge of the function of *PARD3B* is insufficient to categorise the mutations identified in our cohort as being possibly causal.

2.3.8 A *de novo* deletion that overlaps with *OFD1* is highly likely to be causal

One of the candidate CNVs is the *de novo* 21 kb deletion g.13770686_13791294del on Xp22.2 found in F14, a female fetus with ventriculomegaly and agenesis of the corpus callosum. The breakpoint positions given here are approximate. The deleted region covers most of the gene *OFD1* (MIM 300170), 15 probe regions, and has a CoNVex score of 26 (Figure 2-6A). Mutations in *OFD1* cause orofaciodigital syndrome 1 (MIM 311200), which causes malformations of the mouth, face, and digits, and in 40% of cases central nervous system involvement, including absence of the corpus callosum (105). This deletion is highly likely to be causal on the basis of this high degree of overlap between the phenotype of F14 and the known phenotype caused by *OFD1* mutations. The mutation has been confirmed by aCGH and the results returned to the family. This is excellent news for the family as the risk of recurrence is very low at <1%, and would only recur in the unlikely event of gonadal mosaicism.

2.3.9 Inherited recessive or X-linked SNVs in five fetuses are possibly causal, in the preliminary round of analysis

Inherited variants in five of the fetuses are possibly causal. These variants have been verified by Sanger sequencing of whole genome amplified genomic DNA. These variants were identified during the preliminary round of analysis of inherited variants, and do not all remain ‘possibly causal’ candidates following the final round of analysis.

In F5 who had cardiac truncus arteriosus, type B interruption of the aortic arch and pyloric stenosis, I found the compound heterozygous variants c.2189G>A (p.730R>Q) and c.721C>G (p.241P>A) in *DLC1* (MIM 604258, ENST00000276297). Homozygous *DLC1* knockout mice are embryonic lethal with deformities of brain and heart (106).

In F6, whose laterality phenotype has been described, I found the compound heterozygous variants c.4264G>A (p.1422V>M) and c.3686G>A (p.1229R>Q) in *RERE* (MIM 605226*,* ENST00000337907). *RERE*, which is in the retinoic acid pathway, has a role in establishing bilateral symmetry. Although it is not a known human disease-associated gene, homozygous knockout mice develop asymmetrically and have cardiovascular outflow defects. Homozygous zebrafish mutants have cartilage and skeletal defects, abnormal fins and otoliths, reduced viability, deformed brains, and absent gills (107-109). In total I have identified two genes with *de novo* mutations and one gene with inherited variants that could possibly account for the phenotype in F6. It is not possible to say which is most likely to be causative, as none of the candidate genes are known to harbour variants that cause the exact phenotype reported here. One possibility is that multiple variants contribute to this multisystemic phenotype, as has been reported in other exome sequencing studies of rare disease (3, 11).

In F8, with a complex cardiac anomaly on ultrasound including transposition of the great arteries, we found the compound heterozygous variants c.1208_1210delGAG (p.G404del) and c.14194A>G (p.4732K>E) in *RNF213* (MIM 613768, ENST00000582970). *RNF213* has a possible role in vascular development, has been implicated in moyamoya disease, and zebrafish knockdowns have abnormal blood vessels (110).

In F12, ultrasound demonstrated significant ventriculomegaly and unilateral talipes. The homozygous in-frame deletion c.244_249delGGCGGC (p.G82_G83del) in *DACH1* (MIM 603803, ENST00000305425) was identified. *DACH1* is involved in the development of various structures including the limbs and nervous system, and homozygous knockout mice die shortly after birth (111-113).

Finally, F13 had multiple abnormalities including a multicystic-dysplastic kidney, distorted ribs and spine, brain defects and bilateral talipes equinovarus. Here I discovered the compound heterozygous missense variants c.1918C>T (p.640R>C) and c.5205C>A (p.1735H>Q) in *FRAS1* (MIM 607830, ENST00000264895). *FRAS1* variants can cause Fraser syndrome (MIM 219000), severe cases of which include kidney abnormalities such as cysts (114). FRAS1 has a role in renal development and epidermal adhesion (115). Additionally, *FRAS1* transcripts are upregulated in polycystic mouse kidneys (116), and knockout mice have severely defective kidney development, along with syndactyly (117). Homozygous zebrafish mutants have malformed fins and pharyngeal pouches, suggesting a possible role for FRAS1 in skeletal development (118, 119).

2.3.10 The variant prioritisation program eXtasy identifies 36 possibly causal variants, with an enrichment of *de novo* mutations

While manual variant prioritisation using a decision tree is a thorough and nuanced approach, it is neither objective, nor suitable for much larger cohort sizes. Therefore, I decided to investigate two computational methods of variant prioritisation: eXtasy and PhenoDigm. The first aim of this was to assess the utility of these programs in comparison to manual methods, with a view to developing recommendations for larger cohorts. My second aim was to identify any interesting candidate genes from this cohort that my manual method missed.

eXtasy uses a statistical learning approach to prioritise candidate non-synonymous SNVs, taking into account the phenotype of the individual (120). The input to eXtasy is the merged VCF files of the proband, and a list of phenotypes of the proband encoded as human phenotype ontology (HPO) terms (75). Essentially, eXtasy looks at many different features of other genes in which variants are known to cause the phenotype of interest. These features include the haploinsufficiency score of the gene, multiple estimates of the variant impact including PolyPhen, SIFT, and Mutation Taster scores, and multiple estimates of the level of conservation of the genomic region. Next, eXtasy calculates these features for each candidate non-synonymous SNV in the individual. Finally, a random forest algorithm is used to compute an ‘eXtasy score’ for each SNV for each phenotype, which lies between 0 and 1, and is a measure of the probability that each SNV causes each phenotype. The higher the similarity between the features of the variant in the individual, and the features of variants known to cause the phenotype, the higher the eXtasy score will be. An eXtasy score of >0.5 is considered indicative that the variant warrants further investigation. If no genes are known to be associated with a given phenotype, eXtasy will not be able to compute that phenotype.

Next, eXtasy computes a combined p-value that indicates, for each non-synonymous SNV, the significance level, merged across all phenotypes of the individual. There are typically around 9000 non-synonymous SNVs per individual, so a stringent Bonferroni-corrected p-value threshold of significance of 5.6×10^{-6} is probably appropriate. If a combined p-value cannot be calculated (for example because there are not enough phenotypes), the highest eXtasy score for a SNV is an alternative metric by which to rank them. However, where available, the p-value is preferred, because although there may be a high score for an individual phenotype, this does not necessarily equate to a high overall score, if there are lots of additional phenotypes for that patient with a low score. For this experiment, all candidate genes with a maximum eXtasy score >0.5 also have a combined p-value of $< 5.6 \times 10^{-6}$.

There are 475 candidate non-synonymous SNVs in this cohort, 25 of which are *de novo* (Table 2-2, not including those in F31-F33, which were sequenced subsequent to these analyses), and 450 of which are inherited recessive or X-linked (Appendix 2). Of these 475, 36 (in 24 genes) have a significant likelihood of causing the phenotypes, according to eXtasy ($p < 5.6 \times 10^{-6}$) (Table 2-4).

Two of the three mutations I classified as highly likely to be causal are non-synonymous SNVs. Both of these (in *COL2A1* in F20 and in *FGFR3* in F23) were identified as likely candidates in eXtasy. Eight of the eleven variants I classified as possibly causal are non-synonymous SNVs. Three of these (in *NF1* in F6 and two in *RERE* in F6) were identified as likely candidates in eXtasy.

Only 5.3% of the 475 candidate non-synonymous SNVs are *de novo*, but of the 36 that were identified as likely candidates in eXtasy, 6 (16.7%) are *de novo*. This represents a significant enrichment of *de novo* mutations in the variants identified by eXtasy (p = 0.016, Fisher's exact test). This is very interesting given that *de novo* mutations are particularly likely to cause rare disease (11, 55, 57), and that eXtasy is blind to the mode of inheritance of the candidate variants.

ID	CHR	POS	REF	ALT	Gene	COMBI P	MAX eXtasy	Variant type
F1	8	101718965	G	A	*PABPC1*	2.14E-16	0.4	inherited
F1	8	101718968	C	T	*PABPC1*	3.30E-12	0.376	inherited
F1	8	101719138	C	T	*PABPC1*	1.00E-14	0.396	inherited
F1	8	101719201	A	G	*PABPC1*	1.19E-11	0.41	inherited
F2	16	9857047	G	A	*GRIN2A*	1.20E-12	0.292	*de novo*
F2	19	49113215	G	A	*FAM83E*	4.40E-07	0.128	inherited
F5	2	179634421	T	G	*TTN*	1.62E-06	0.36	inherited
F6	1	8418331	C	T	*RERE*	1.64E-07	0.376	inherited
F6	1	8418909	C	T	*RERE*	2.27E-06	0.284	inherited
F6	7	103141235	G	A	*RELN*	5.12E-09	0.286	inherited
F6	7	103205827	G	C	*RELN*	7.10E-13	0.46	inherited
F6	17	29562669	G	A	*NF1*	1.04E-17	0.624	*de novo*
F6	19	41754430	G	A	*AXL*	6.28E-12	0.614	inherited
F9	20	61288233	G	A	*SLCO4A1*	4.96E-09	0.292	inherited
F10	1	39851427	G	A	*MACF1*	2.78E-11	0.644	inherited
F10	1	39901245	A	G	*MACF1*	1.70E-14	0.714	inherited
F10	8	20069263	G	T	*ATP6V1B2*	9.96E-22	0.49	*de novo*
F10	9	91994007	C	T	*SEMA4D*	4.99E-08	0.18	*de novo*
F11	X	30322699	T	C	*NR0B1*	2.41E-07	0.24	inherited
F13	2	1459885	A	G	*TPO*	8.57E-14	0.24	inherited
F13	2	1544464	C	T	*TPO*	2.53E-19	0.388	inherited
F17	1	68960131	T	C	*DEPDC1*	1.34E-11	0.308	inherited
F17	1	68960186	T	C	*DEPDC1*	2.54E-07	0.162	inherited
F18	2	179611552	C	T	*TTN*	4.29E-08	0.672	inherited
F18	3	135969390	A	C	*PCCB*	2.08E-12	0.632	inherited
F18	3	136019898	C	T	*PCCB*	1.09E-11	0.458	inherited
F18	X	138644189	C	T	*F9*	2.46E-10	0.458	inherited
F19	16	87723683	G	A	*JPH3*	5.03E-06	0.454	inherited
F20	12	48369853	C	A	*COL2A1*	2.24E-06	0.654	*de novo*
F21	6	51656129	C	G	*PKHD1*	1.67E-07	0.714	inherited
F21	6	51768399	A	T	*PKHD1*	3.59E-09	0.888	inherited
F23	2	179610967	C	T	*TTN*	3.89E-18	0.626	inherited
F23	4	1806099	A	G	*FGFR3*	2.71E-28	0.902	*de novo*
F23	11	70336479	C	T	*SHANK2*	2.10E-10	0.384	inherited
F23	15	22969250	C	T	*CYFIP1*	3.04E-14	0.718	inherited
F23	X	19398315	C	T	*MAP3K15*	1.57E-12	0.268	inherited

Table 2 - 4: Candidate genes identified as possible causal by eXtasy.
Table contains genes with eXtasy combined p value < 5.6 x 10^{-6}. COMBI_P = combined p value. MAX eXtasy = maximum eXtasy score across the phenotypes. These are both measures of how likely a variant is to cause the fetuses phenotypes.

2.3.11 The variant prioritisation program PhenoDigm identifies possibly causal variants in 18 genes

The PhenoDigm program identified possibly causal disease-associated genes on the basis of overlap between the phenotype of a patient, and the mouse phenotype caused by knocking out the orthologue of genes in which variants have been found in the patient (121). If no mouse model has been generated and phenotyped for a gene of interest, PhenoDigm cannot be used. Around 32% of mouse protein-coding genes have a phenotyped model available (personal communication from Dr Damian Smedley).

The input to PhenoDigm is a list of candidate genes, and a list of phenotypes encoded as HPO terms, for each patient. The output is, for each candidate gene, two scores indicating the degree of overlap of each patient phenotype with the mouse model. These scores are the Information Content (IC) and Jaccard Index (simJ) scores. If the geometric mean of these two scores is >1.5, variants in that gene are possibly causal. However, as for eXtasy, there may be considerable overlap for one HPO term, but this does not necessarily mean there is high *overall* overlap across all phenotypes observed in the patient. The version of PhenoDigm that was used for these analyses was an early version that used only mouse phenotype data, whereas more recent versions incorporate data from zebrafish.

There are 390 candidate genes in this cohort (where a gene recurs in multiple fetuses, I have counted it that number of times here): 31 have *de novo* mutations (Table 2-2, not including those in F31-F33, which were sequenced subsequent to these analyses), 7 are in CNVs (Table 2-3), and 352 have inherited recessive or X-linked variants (Appendix 2). Of these 390, 99 have a phenotyped mouse model, and of these, 18 are possibly causal disease-associated genes identified by PhenoDigm (Table 2-5).

ID	Gene	Fetus HPO term	Model MPO term	Geo Mean	Variant type
F3	*NCOR2*	Ventricular septal defect	Ventricular septal defect	2.23	Inherited
F5	*TTN*	Ventricular septal defect	Heart left ventricle hypertrophy	1.83	Inherited
F6	*FOXC1*	Ventricular septal defect	Ventricular septal defect	2.23	Inherited
F6	*NF1*	Double outlet right ventricle	Persistent truncus arteriosus	2.28	*De novo*
F6	*TGIF1*	Abdominal situs inversus	situs inversus	2.66	Inherited
F7	*TTN*	Ventricular septal defect	Heart left ventricle hypertrophy	1.83	Inherited
F9	*GNAS*	Abnormality of the thymus	Thymus atrophy	2.17	Inherited
F13	*FRAS1*	Talipes	Clubfoot	2.41	Inherited
F13	*PTCH1*	Missing ribs	Decreased rib number	2.61	Inherited
F13	*TGIF1*	Microcephaly	Microcephaly	2.17	Inherited
F17	*ABCA3*	Pulmonary hypoplasia	Increased wet-to-dry lung weight ratio	2.19	Inherited
F19	*DNAH5*	Defect in the atrial septum	Ostium secundum atrial septal defect	2.36	Inherited
F19	*NCOR2*	Defect in the atrial septum	Ventricular septal defect	1.96	Inherited
F20	*COL2A1*	Abnormality of the lower limb	Short femur	1.83	*De novo*
F20	*SMPD1*	Choroid plexus cyst	Abnormal choroid plexus morphology	2.55	Inherited
F23	*FGFR3*	Short ribs	Short ribs	2.60	*De novo*
F25	*HIF3A*	Pleural effusion	Abnormal pulmonary artery morphology	1.61	Inherited
F29	*TTN*	Tricuspid regurgitation	Increased left ventricle diastolic pressure	1.63	Inherited

Table 2 - 5: Candidate genes identified as possibly causal by PhenoDigm.
Table contains genes with Geo mean >1.5. For each gene, only the phenotype with the highest Geo mean is shown. Geo mean = geometric mean of the SimJ and IC scores; HPO = human phenotype ontology; MPO = mammalian phenotype ontology.

PhenoDigm identified *COL2A1* in F20 and *FGFR3* in F23 as containing possibly causal variants. These were also identified by the decision tree method, and by eXtasy. Two of the eight genes containing variants classified as possibly causal by the decision tree method were also identified as likely candidates using PhenoDigm (*NF1* in F6 and *FRAS1* in F13). *NF1* was also prioritised by eXtasy.

2.3.12 There is a degree of overlap between the variants identified as possibly causal by the three different prioritisation methods

The variants that were prioritised by the three different variant prioritisation methods (manual decision tree, eXtasy and PhenoDigm) overlap somewhat (Figure 2-9). All three methods prioritised *FGFR3*, *COL2A1* and *NF1*. Both the decision tree and eXtasy prioritised *RERE*. Both the decision tree and PhenoDigm prioritised *FRAS1*. Both eXtasy and PhenoDigm prioritised *TTN*. I did not prioritise *TTN* manually because it is an exceptionally large gene in which many variants fall by chance. Additionally, there are five prioritisations unique to the decision tree, 19 unique to eXtasy and nine unique to PhenoDigm.

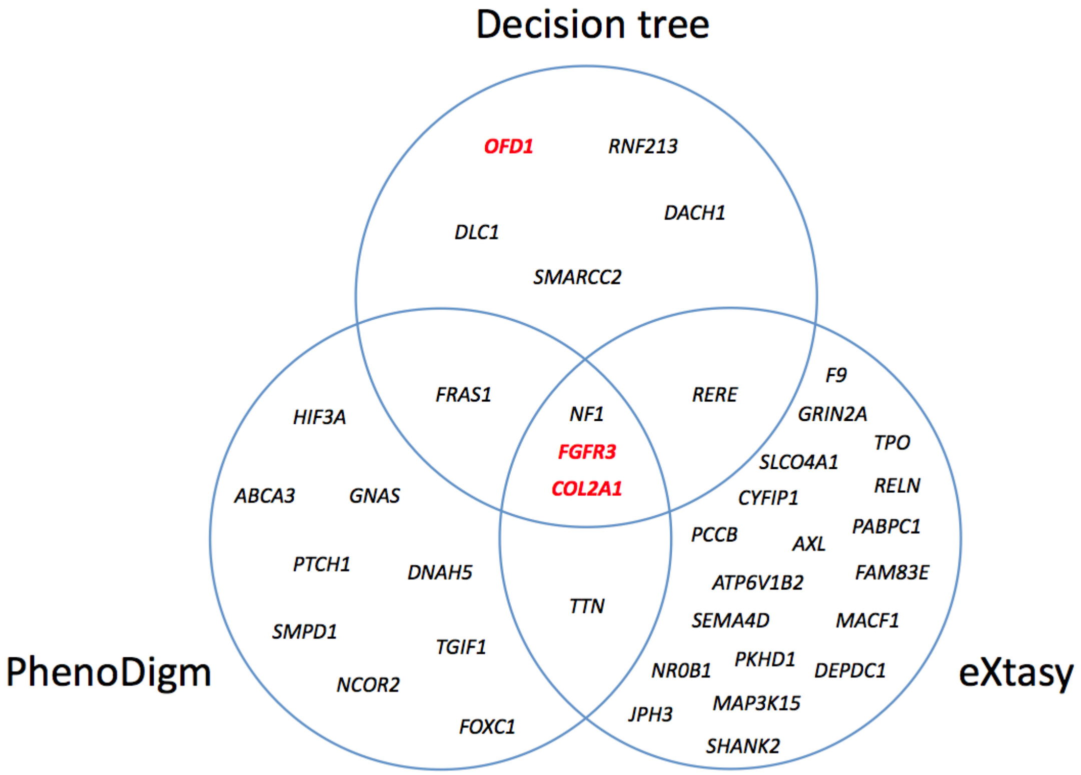

Figure 2 - 9: Venn diagram showing overlap between the genes prioritised by each of the three methods.
The genes named in the decision tree circle include both 'highly likely to be causal' and 'possibly causal' candidates, with the former in red.

It is important to note that, while this overlap is interesting, the results are not strictly speaking directly comparable, because some methods are not capable of identifying the same candidates as others. For example, eXtasy could not have identified *OFD1* as a candidate, because the variant in this case is a deletion and eXtasy only interrogates non-synonymous SNVs. Similarly, PhenoDigm could not have identified *SMARCC2* as a candidate, because a mouse model of this gene is not available.

2.3.13 The continuing need for manual curation

I further investigated the variants prioritised by eXtasy and PhenoDigm, in order to decide whether I should consider upgrading any to my ‘possibly causal’ or ‘highly likely to be causal’ categories. For eXtasy, I concluded that most of the additional variants that it prioritised should not be upgraded because they either had no obvious link to the fetal phenotype, recurred in multiple cases with non-overlapping phenotypes, or were found in a fetus for which I had found a clearly causal variant. However, on further investigation I decided that one of the genes highlighted by eXtasy should be upgraded: *MACF1* in F10, which is discussed further below. The PhenoDigm results did not lead me to upgrade any variants because they all either recurred in multiple cases with non-overlapping phenotypes, or only had overlap with a small proportion of the fetal phenotypes. This emphasises the continuing need for manual curation of results of computational gene prioritisation methods.

2.3.14 Inherited recessive or X-linked SNVs in five fetuses are possibly causal, in the final round of analysis

As I have explained, I reanalysed the inherited recessive or X-linked variants using a slightly more sensitive and specific filtering protocol, incorporating the additional samples F31-F33, and upgrading *MACF1* in F10 to a ‘possibly causal’ gene on the basis of the eXtasy analysis. For this final round of analysis, I detected a mean of 21,444 high-quality coding SNVs and indels per individual (Table 2-1). Filtering for rare, functional variants leaves a mean of 5.3 candidate genes per fetus (range of 0-15) with a total of 139 different candidate genes across the 30 fetuses, containing 269 rare functional variants. Of these variants, 262 are missense, four are frameshift, and three are nonsense (Appendix 3).

Inherited variants in five of the fetuses are possibly causal, in this final round of analysis. These variants have been verified by Sanger sequencing of whole genome amplified genomic DNA. The possibly causal variants in *DLC1* in F5, *RERE* in F6, and *FRAS1* in F13, are as I described in section 2.3.9. However, I now also consider *PRKDC* variants in F1 and *MACF1* variants in F10 as possibly causal, and I no longer consider *RNF213* variants in F8 or *DACH1* variants in F12 to be possibly causal.

In F1, a male fetus with multiple abnormalities including limb defects, craniofacial defects, anogenital defects, heart defects, a tracheal oesophageal fistula and renal agenesis, I found the compound heterozygous variants c.9598C>T (p.3200P>S) and c.1420G>T (p.474V>F) in *PRKDC* (MIM 600899, ENST00000338368). *PRKDC* encodes DNA-PKcs, which, in complex with Ku, is required for the DNA double-strand break repair mechanism non-homologous end joining. In humans, *PRKDC* variants can cause severe combined immunodeficiency due to defective V(D)J recombination, and severe cases can also have abnormalities of the brain, face, limbs, and anogenital organs (122). *PRKDC* was not identified as a candidate gene in the preliminary round of analysis because the study described here was published in July 2013, subsequent to the preliminary analysis.

F10 had fetal akinesia syndrome probably caused by neuroaxonal dystrophy. I found the compound heterozygous variants c.5323G>A (p.1775E>K) and c.8626A>G (p.2876I>V) in *MACF1* (MIM 608271, ENST00000372925), which encodes cytoskeletal protein microtubule-actin cross-linking factor 1. Knockout of the mouse orthologue causes defects in axonal extension (123). This was not a candidate in the preliminary round of analysis because it was brought to my attention by the eXtasy variant prioritisation.

DACH1 variants in F12 and *RNF213* variants in F8 were considered highly likely to be causal after the preliminary round of analysis, but not after the final round. This is because for the final round I added a new minor allele frequency filter (<0.01 in an internal control cohort of 2172 individuals). The *DACH1* variant in F21 had a frequency in the control cohort of 0.47. One of the compound heterozygous variants in *RNF213* in F8 had a frequency in the control cohort of 0.014. It is therefore highly likely that these variants do not cause the structural abnormalities in these fetuses.

F19 has a high number of inherited, apparently rare variants (Appendix 3). F19 is of Indian ancestry, whereas the majority of the cohort is of European ancestry. It is likely therefore that some of the apparently rare variants that I have identified in F19 are in fact more common in this population, but I have not been able to identify them as such due to an underrepresentation of individuals of Indian ancestry in the databases I used to filter the variants.

2.3.15 The estimated diagnostic yield of this study is 10%

According to the classification system described, and in close collaboration with the clinical team at the University of Birmingham, I identified three mutations that are highly likely to be causal: the *de novo* mutation in *FGFR3* in F23, the *de novo* mutation in *COL2A1* in F20 and the *de novo* deletion covering *OFD1* in F14. Additionally, I identified seven variants (in five additional fetuses) that are possibly causal: two *de novo* and five inherited. Candidate genes in all categories are summarised in Table 2-6. Out of our cohort of 30, this represents a minimum diagnostic yield of 10%, although due to the relatively small size of the cohort, this estimate of 10% has a broad 95% confidence interval of 3.5% - 25.6%.

ID	Sex	De *novo*	Inherited autosomal recessive (comp het)	Inherited autosomal recessive (homozygous)	Inherited X-linked	CNV
F1	M	.	*HEPHL1;* ***PRKDC****; ZNF44*	.	*BCORL1; FAM47A; KCNE1L; MAGEA6; ZCCHC12*	.
F2	F	*GRIN2A*	*FAM83E; KIAA1239; KIAA1755; LAMA5; MIA3*	.	.	.
F3[1]	M	*PPFIBP2*	*C16orf91; C9orf79; CCDC144NL; NHSL1*	.	*CCDC22; SHROOM2*	*[H2BFM; H2BFWT]*
F5	M	.	***DLC1****; TTN*	.	*FAM70A; FTHL17; GPR112; PCDH19; RBMXL3; WDR44*	.
F6	F	*C11orf41;* ***NF1****;* ***SMARCC2****; ZHX3*[3]	*FAM188B; RELN;* ***RERE***	*AXL*	.	.
F7	F	*UNC80; WFDC8*	*MUC16; TSC22D1; TTN*	.	.	.
F8	M	*CD244*	*LY75-CD302; TTN; WDR59*	.	*PLXNB3; RBBP7; SRPX2*	.
F9	M	*PARD3B*	*ABCA13; COL6A6; GNAS; KIAA1462; MUC17; SRRM2; TRPM8*	.	*ATP2B3; CCDC22*	.
F10	F	*ATP6V1B2; SEMA4D*	*C19orf28; CDHR1; DNAH10;* ***MACF1***	.	.	.
F11	M	.	*REST*	.	*CITED1; MXRA5; NR0B1*	.
F12	F	.	*FRG1B; TTN; ZNF451*	.	.	.
F13	M	.	***FRAS1****; SPTBN5; TPO*	.	*ALG13; DDX26B; MAP7D3; TLR7*	.
F14	F	*KCTD8; STX12*	*ADNP; ANO7; CENPF; TDRD6*	.	.	*[GPM6B;* ***OFD1****]*
F15	F	*DOCK1*	*ABLIM3; VCAN*	.	.	.
F16[1]	M	*PPFIBP2*	*C16orf91; C9orf79; CCDC144NL; NHSL1*	.	*SHROOM2*	.
F17	F	.	*ABCA3; AKAP11; DEPDC1; PAFAH2; POM121C*	.	.	.
F18	M	*ABCB9;*	*PCCB; TTN;*	.	*CXorf57;*	.

		FAM3D	*ZFHX3*		*DUSP21; F9; FOXR2; HS6ST2; NKAP; RBMX2*	
F19[4]	M	*DNAJC13*[3]*; NLRP1; PARD3B*	*AHNAK2; C20orf90; CD163L1; DNAH1; DNAH5; DNAH6; FSTL4; PHLPP2*	*ADAD2; PCNT*	*COL4A6; GYG2; PNMA3; SATL1; SHROOM2*	*[SSX3; SSX4; SSX4B]*
F20	M	***COL2A1***	*CHD7; EPB41L2; GPR98; VPS13D*	.	*FAM58A; MTCP1NB; PLXNA3; SLC10A3*	.
F21	M	.	*CACNA1H; PKHD1*	*KIF26A*	*ARMCX2; EDA2R; HTATSF1; MAP7D3; MTMR8; MXRA5*	.
F22	M	*TACR2*	*DECR1; DUOXA1; NEB; VPS13C*	*PCDHB7*	*MAP7D3*	.
F23	M	***FGFR3***	*C1orf129; SHANK2; TTN*	*GFM2*	*MAP3K15; MAP7D3*	.
F25	M	*PNLIPRP1; SMARCC1*	*HSPG2; IQGAP3*	.	*BCOR; RAB40A; USP26*	.
F26	M	*KDM5B; STAU2*	*GNRHR2*	.	*HTATSF1; MTMR1; PIR*	.
F27[2]	F	*C2orf40; INSC*	.	.	.	.
F28	F	*PPP6R1*	*CYP24A1; KIAA1109; KIAA1609; SLC39A11*	.	.	.
F29	F	.	*ABCA13; MCF2L2; NLRP12; POM121C; TTN; ZNF831*	*TTN*	.	.
F31	F	*FMNL3*	*FAH*	.	.	.
F32	F	.	.	.	.	.
F33[2]	F	*SEC31B; EGFL6*[3]	*AGRN; NUDT19*	.	.	.

Table 2 - 6: Summary of all candidate genes identified in 30 fetuses with structural abnormalities.

Column headers indicate the type of variant associated with the candidate genes. Bold red text indicates variants that are highly likely to be causal. Bold orange text indicates variants that are possibly causal. Square brackets contain genes in a single CNV. [1]Monozygotic twins; [2]Siblings; [3]Synonymous *de novo* mutation; [4]Indian ancestry.

2.4 Discussion

2.4.1 Summary

In this study, I analysed exome data from 30 parent-fetus trios with a range of fetal structural abnormalities detected from prenatal ultrasound. I identified rare, LOF or functional, *de novo* and inherited (X-linked or recessive) variants. I used a decision tree to interpret the variants, and together with colleagues decide which were likely to be causal. I found a degree of overlap between the genes I classified as causal using this subjective method, and genes prioritised by two different pieces of gene prioritisation software. For three fetuses (10%) I found mutations that were highly likely to be causal. For a further five fetuses (17%), I found variants that were possibly causal. This study is the largest published cohort of fetuses with structural abnormalities to have been exome sequenced to date, and suggests that exome sequencing is a viable diagnostic strategy in these cases.

2.4.2 The diagnostic yield in context

The diagnostic yield of this study was 10%. The typical diagnostic yield of microarrays in cohorts of fetuses with structural abnormalities is 6-10% (22, 34, 40). Only one of the causal mutations identified in this study was a CNV detected by microarray, which highlights the additional utility of exome sequencing, and demonstrates that the detection rate is increased over that achieved by karyotyping and microarrays alone.

Nevertheless, our diagnostic rate is lower than that found in exome sequencing studies of rare postnatal diseases, which is typically around 25% (3, 11, 61). There are several possible reasons for this. First, our estimate of 10%, being based on a relatively small sample size, has a broad confidence interval of 3.5% - 25.6%, meaning that a diagnostic rate of up to 25% could be possible in prenatal samples, and the diagnostic rate in this study might just be lower just by chance. Second, it is likely that in some cases, variants in the same gene will have different phenotypic manifestations between prenatal and postnatal stages of development (124). It seems likely for example, that, for a given variant or gene, one might observe more severe phenotypes *in utero*, which may not be compatible with life postnatally. Given that I interpreted the data in this study by comparing fetal phenotypes to available data, the vast majority of which is postnatal, this makes interpretation more difficult. Similarly, for some of the fetuses in this study the only phenotypic data came from ultrasound scans. There are many phenotypes that cannot be identified from an ultrasound scan including subtle morphological abnormalities, most metabolic phenotypes, and behavioural and cognitive deficits. This potentially incomplete phenotype data also complicates variant interpretation.

In this study, we did not identify any novel disease-associated genes. This is unsurprising because the study is underpowered for this task because of the small cohort size, and variation in phenotypes. However, the recurrence of *de novo* mutations in *PARD3B* in two fetuses with non-overlapping phenotypes is intriguing. The probability of this happening by chance is small ($p = 3.1 \times 10^{-6}$, which does not quite reach the stringent Bonferroni-corrected significance threshold of $p = 2.5 \times 10^{-6}$, but is clearly close to it). Further work such as sequencing of *PARD3B* in larger cohorts of fetuses, or investigation of *PARD3B* function using model organisms, would shed more light on whether these mutations have a role in the phenotypes of these fetuses.

2.4.3 Comparison of variant interpretation methods

I interpreted the variants in this study using three methods: a decision tree, eXtasy and PhenoDigm. Each had advantages and disadvantages. The advantages of using a decision tree include the fact that it is thorough and wide-ranging. I was able to incorporate information from lots of different sources, not all of which are accessible to computational methods. For example, I could search the PubMed literature for studies about each gene. Computationally, this is a difficult task. While text-mining programs have improved greatly in recent years, they are still subject to technical limitations. Also, I could put different weights on different types of information, taking into account what I know about the biology of the phenotype. Again, this is something that would potentially be difficult to automate. For example, typically if a phenotype of an animal model and a human patient with variants in orthologous genes overlap, this strongly suggests that the variants might be causal in the patient. However, if a zebrafish model of a candidate gene found in a fetus with growth restriction had reduced body size, I would not necessarily think this is relevant, because I know that growth delay is a common, fairly non-specific phenotype in zebrafish disease models.

However, the decision tree method has two important disadvantages. First, the very flexibility that I have described leaves room for unconscious bias. I tried to limit this by taking a systematic approach, but there is no escaping the fact that it is a subjective method. For example, one distinction between my ‘highly likely to be causal’ and ‘possibly causal’ categories relies on whether phenotypes overlap ‘to a high degree’, or ‘somewhat’, respectively. There is no quantitative distinction between these groups. Second, it is a labour-intensive method. I estimate that it took me roughly 2-3 hours to categorise the candidate genes for each trio, depending on the number of candidates, and the amount of information available for those candidates. This method was therefore feasible for 30 trios, but would be out of the question for 1000 trios, and probably too slow even for 100 trios. This is why I additionally investigated two computational methods, both of which solve both of these problems.

The variants categorised as interesting by eXtasy had some overlap with those I highlighted using the decision tree. Additionally, they were significantly enriched for *de novo* mutations. In particular, the results from eXtasy highlighted the possibility that *MACF1* variants in F10 are possibly casual. While it is unsurprising that eXtasy prioritises known genes because it is trained on known disease-associated genes, these observations do emphasise the potential of eXtasy as a gene prioritisation tool, and highlight its potential for novel disease-associated gene discovery. However, there are several limitations to the program, too. Currently, it can only be used to prioritise non-synonymous SNVs. Also, it requires information on known genetic causes of the phenotypes of interest. If there are no known genetic causes of an observed phenotype, then the program cannot be used. Finally, it is not always obvious why eXtasy has prioritised a particular variant, when it is in a gene with no obvious link to phenotype. Clearly, the gene has some similarity to another gene known to cause the phenotype. However, the information about what that other gene is, and in what way it is similar, is not easy to extract. Therefore these cases are very difficult to interpret.

Of the ten variants that I initially classified as highly likely to be causal or possibly causal, PhenoDigm also highlighted four of them as interesting. This is a promising degree of overlap. The main disadvantage of PhenoDigm is that if there is no animal model for a particular gene, it cannot be used. This limits its utility in practice, and means that it could not be used as the sole method of variant prioritisation, at least until a higher proportion of mouse genes have phenotyped knockouts. Similarly, there are cases where the phenotype of a human and the phenotype of a mouse with a variant on the orthologue of the same gene are not similar (125). While these cases are not typical, they could lead to misleading results from PhenoDigm. The other disadvantage of PhenoDigm is that it can give a very significant score for a gene when there is overlap of a single phenotype. But this does not equate to a high degree of overall phenotypic overlap. For example, PhenoDigm identified *ABCA3* as an interesting candidate in F17, because the mouse has a similar lung defect to the fetus. However, the fetus had 14 phenotypes, only two of which overlapped with the mouse. Almost all of the phenotypes of the mouse model had to do with the lungs, whereas the fetus had many additional affected systems that did not recapitulate in the mouse model. Therefore, I concluded that the *ABCA3* variants are unlikely to be causal.

From my comparison of these three methods, I concluded that each of the computational tools identified most of the same high priority candidates that the manual method did. However, they each have technical limitations. Furthermore, they currently have insufficient sensitivity and specificity to replace manual investigation by a researcher. For a large-scale exome sequencing project, my recommendation for a variant prioritisation approach, based on my experience described here, would be to employ at least two computational approaches of gene prioritisation. Where the results overlap, it is likely that those candidate genes are strong candidates, assuming that the programs take reasonably independent approaches, and assuming that huge genes such as *TTN*, which are often problematic in such approaches are considered separately. Candidates identified by one program but not the other should undergo manual curation by a researcher to decide whether they are likely to be causative. Finally, the technical limitations of the programs must be overcome. For example, eXtasy only prioritises non-synonymous SNVs, so all other categories of variants would have to be considered separately. In addition to this, it is necessary to use robust statistical assessment to determine whether the candidate variants were likely to have arisen by chance.

2.4.4 The ethics of next generation sequencing for prenatal genetic diagnosis

The many thorny ethical issues surrounding NGS in the clinical context have been extensively debated, chief among them are whether to report incidental findings, and how to report variants of unknown significance (VOUS) (126). In the prenatal context, the issues are similar but amplified, partly due to the possibility of termination of the pregnancy. One possible application of prenatal sequencing that raises some unique ethical questions is widespread use in the general population.

In some cases, widespread use of prenatal sequencing in the general population could identify a pathogenic variant that causes a severe, distressing, and lethal phenotype and is highly penetrant, at an earlier stage than an ultrasound scan could have found structural abnormalities. An example of such a variant might be missense changes in *FGFR3* that cause thanatophoric dysplasia (23). In these scenario, earlier detection would undoubtedly be better for families. It would avoid potentially devastating news later in pregnancy, in the neonatal period, or even later in childhood. If the families elect to terminate the pregnancy, distress is generally less severe at an early stage of pregnancy. For families who choose to continue with the pregnancy, early diagnosis may offer a more accurate prognosis, more time to prepare, and in some cases the option to start treatments earlier. Therefore, such families would definitely benefit from prenatal sequencing.

However, in other, less clear-cut cases, the disadvantages of widespread use of prenatal sequencing in the general population may outweigh the advantages. Identification of VOUS is virtually inevitable during prenatal sequencing. For example, a predicted pathogenic variant may be identified in a known developmental disorder gene, but if it has never been reported before it may be very difficult to accurately predict the phenotype. The ethical issues of returning VOUS to families have been considered in the context of CNVs discovered by aCGH. Some research suggests that receiving information on VOUS during pregnancy can be very distressing (127). Therefore, some researchers and clinicians think that they should not be reported to families, and that their detection should be limited in the first place by using targeted tests (37). Others think that it is paternalistic to withhold this information (128). If VOUS were to be returned, it is imperative that families receive extensive genetic counselling before and after prenatal sequencing. These issues are still under debate, and it is important for clinicians and researchers to come to a consensus on the issue of reporting VOUS, prior to any widespread use of prenatal exome sequencing in the general population, because interpreting variants identified by exome sequencing is generally more difficult than those identified by aCGH, and there will be a higher number of VOUS identified.

Another question is whether return to families information on variants that are likely to cause late-onset disease, or have incomplete penetrance, such as a *BRCA1* variant that confers an 80% risk of developing cancer later in life (129). Some argue that families have a right to this information to do with what they will, even if it will result in increased termination rates, and termination of some healthy fetuses (130). An alternative is to do more targeted sequencing based on the indication for the test, so as to avoid incidental findings.

There are currently more questions than answers regarding the ethics of widespread implementation of prenatal exome sequencing in the general population. Nevertheless, many pertinent issues have already been thoroughly discussed in the context of postnatal clinical sequencing, or interpretation of prenatal aCGH results. While prenatal exome sequencing clearly poses additional specific ethical challenges, it is likely that with continued open debate amongst clinicians and researchers, along with sensitive and thorough genetic counselling to families, these can be overcome.

2.4.5 Next generation sequencing is the future of prenatal genetic diagnostics

From a scientific perspective, it seems inevitable that NGS is the future of prenatal genetic diagnostics. Nevertheless, many questions remain to be answered before prenatal NGS could become widespread. These include issues of cost effectiveness, clinical utility, ethics, and interpretation of variants.

To address some of these, the Wellcome Trust and the Department of Health in the UK have awarded a Health Innovation Challenge Fund grant to the collaborative Prenatal Assessment of Genomes and Exomes (PAGE) project. This will involve WTSI, the University of Cambridge, the University of Birmingham, Birmingham Women's Foundation Trust, University College London and Great Ormond Street Hospital (London, UK). One thousand fetuses with structural abnormalities, along with maternal and paternal samples, will undergo exome sequencing or whole genome sequencing from invasively sampled material. The results of this study are expected to yield insights into the genetic causes of fetal abnormalities, and pave the way scientifically, clinically, and socially for large-scale implementation of NGS in the UK's prenatal arena. Additionally, the increased size of the PAGE cohort compared to that of this study will increase power to identify novel disease-associated genes, and allow for a more accurate estimate of diagnostic yield.

Exome sequencing is currently considered more cost-efficient than whole-genome sequencing for clinical diagnostic purposes. However, for several reasons, I predict an eventual move towards whole-genome sequencing rather than exome sequencing for clinical diagnostic purposes, including in prenatal samples. First, there are many examples of non-coding variants that can cause congenital abnormalities including pancreatic agenesis and malformations of the digits (131, 132). These variants would usually not be detected by exome sequencing. Second, while the costs of NGS are falling rapidly, if the costs of the exome capture step do not fall in line with this, at some point whole-genome sequencing may become more cost-effective than exome sequencing (133). Third, in exonic regions that are difficult to capture (for example because they are GC-rich), whole genome sequencing actually results in higher sensitivity of variant calling in coding regions than exome sequencing does (63). Finally, a major reason why whole-genome sequencing is currently often avoided is that interpretation of non-coding variants is very difficult. However, with large-scale whole-genome projects being planned, this is also likely to start becoming easier (134).

Another important advance in prenatal diagnostics would be the ability to detect *de novo* mutations non-invasively, by the sequencing of maternal cfDNA. Currently, this requires sequencing to a depth that has not yet been achieved genome-wide. Further technical advances in coming years are likely to render this possible, making this technique far more useful. For example, improvements in calling algorithms could reduce the required depth of coverage to detect *de novo* fetal variants. Another possibility is the development of supremely accurate whole genome amplification methods, which would allow a sufficient quantity of DNA to be obtained from a maternal plasma sample to achieve the required depth. This would also require continuing decreases in sequencing costs, because it would involve generation of a huge amount of data.

In regard to this cohort, I think that the most fruitful next step would be to perform further, functional investigation of some of the ‘possibly causal’ candidate genes. For example, phenotypic investigation of a zebrafish *PARD3B* knockdown embryo might help to clarify the role of this gene in development. Similarly, there are currently no animal models of *SMARCC2*. While this gene may prove to be lethal if completely knocked out because it is a chromatin regulator, a heterozygous mouse or a zebrafish knockdown may be able to clarify whether the *de novo SMARCC2* mutations found in F6 contributes to the phenotype.

In conclusion, the main outcomes of this project are as follows. We have achieved an approximate diagnostic yield of 10% in this small cohort. All of these 10% were *de* *novo* mutations, which would allow families to be counselled as to a low recurrence risk. We found possible genetic causes for an additional 17% of the cohort. While we could not confidently ascribe pathogenicity in these cases, these data might aid variant interpretation for other researchers who might come across candidate pathogenic variants in those genes. More widely, we have demonstrated the utility and efficacy of exome sequencing for the purposes of prenatal genetic diagnostics, and paved the way for the PAGE project to expand upon these findings.

3 Case-control analysis of 565 known and candidate intellectual disability-associated genes

3.1 Introduction

3.1.1 The impact of intellectual disability

Intellectual disability (ID) is diagnosed in patients who have an intelligence quotient (IQ) of below 70, along with problems with adaptive functioning (such as problems communicating or caring for themselves), where these symptoms began before the age of 18 (135). ID is typically classified as mild (IQ 50-70) or severe (IQ below 50) although other categories can be used. It is phenotypically heterogeneous; in addition to variable IQ and different manifestations of problems with adaptive functioning, it often occurs in conjunction with other abnormalities, such as seizures, behavioural difficulties, dysmorphic facial features, or other developmental disorders such as congenital heart disease (CHD). A particularly common comorbidity of ID is autism spectrum disorder (ASD), with 28% of people with ID also suffering from ASD (136). ID with additional comorbidities is often classified as 'syndromic', and cases with no additional symptoms are 'non-syndromic' (137). However, recent opinion in the ID research community has shifted away this dichotomous categorisation, in favour of considering ID as a spectrum, with variable additional phenotypes. This is partly because subtle comorbidities or specific intellectual disabilities shared among groups of patients are often not obvious until they are retrospectively grouped according to aetiology (135).

Collectively, ID is a very common developmental disorder, with a prevalence of around 1-2%, but estimates of prevalence vary widely depending on factors including the definition of ID, the population studied, and age group (138, 139). Importantly, the prevalence also depends on sex, with males accounting for ~57% of ID cases (140). The majority of patients with ID require extensive medical, financial and personal support throughout their lives, causing ID to be one of the most costly diseases in high-income countries (141). Because ID is so prevalent, it therefore has a profound impact not only on patients and their families, but also on healthcare providers and society as a whole.

The causes of ID are wide ranging and include environmental and genetic factors. Environmental factors that are associated with increased risk of ID include malnutrition during infancy, prenatal exposure to alcohol or the rubella virus, childhood exposure to lead, brain injury during birth, and low birth weight (142-146). Fetal alcohol syndrome affects 0.1-0.7% of births, and is the most common preventable cause of ID in high-income countries. With this exception, environmental factors disproportionately affect people in low-income countries, and explain the increased prevalence of ID in such countries (139).

Genetic causes of ID have been recognised for many decades. Lionel Sharples Penrose was the first to conduct a large study on the subject, which was published in 1938 (147). He assembled and investigated a cohort of 1280 cases of ID. His pioneering observations included the sex bias in prevalence of ID, and the fact that related patients often have similar phenotypes. Historical studies such as this draw attention to two relevant ethical issues. First, ID, possibly more than any other medical condition, uses terminology that has evolved. In Penrose's study, for example, patients are classified according to whether they are "dull", "simpletons", "imbeciles" or "idiots". By 1960, these offensive terms had been replaced in the medical and research communities by the term mental retardation. Gradually, this term too attracted derogatory connotations, and in 2009 a law (known as Rosa's law) was passed in the USA officially replacing it with the term intellectual disability. Second, early studies of the genetics of ID are tainted by their unpleasant association with the eugenics movement. For example, in "The Eugenics Review", Eliot Slater describes aspects of Penrose's study to be "of profound eugenic significance" (148). J. B. S. Haldane, commenting on Penrose's study in Nature, took a moderate approach, emphasising the complexity of the aetiology of ID, and calling the claims that it could be largely eliminated by sterilisation of patients to be "extravagant" (149).

From these beginnings, research into the genetics of ID and intelligence has flourished. Intelligence is a quantitative trait, and is highly heritable (150). Mild, non-syndromic ID represents the bottom of the normal distribution of IQ, and these cases are likely to be influenced by multiple genetic and environmental factors each with a small effect size, as for any quantitative trait. To start to understand the genetic architecture of these cases will require genome-wide association studies with extremely large sample sizes (151). However, moderate to severe ID is thought to be usually caused by a single pathogenic variant with a large effect. Identification of these variants and understanding how they cause ID is of great importance.

3.1.2 Discovery of intellectual disability-associated genes

Cytogenetically visible chromosomal aberrations comprising aneuploidies, rearrangements, and large copy number variants (CNVs) cause around 15% of ID cases (152). Trisomy 21 (also known as Down syndrome) is the most common genetic cause of ID, and was the first ID-associated variant to be discovered. It accounts for ~10% of ID cases (152), and the molecular defect was first identified in 1959, although the syndrome had been recognised since 1866 (153).

The introduction of chromosomal microarrays increased the resolution at which variants could be identified to the submicroscopic level. Submicroscopic CNVs are a frequent cause of ID. For example, heterozygous *de novo* 17q21.31 microdeletions (500-650 kb) can cause a syndrome comprising ID, motor and speech delay, dysmorphic facial features and hypotonia (154). Another study used array comparative genomic hybridisation (aCGH) on a large cohort to demonstrate that submicroscopic CNVs (with a median size of 213 kb) account for ~14% of ID cases (155). Interestingly, they also showed that CNVs disproportionately cause syndromic rather than non-syndromic ID, especially where the additional abnormalities are structural (such as cardiovascular or craniofacial defects). Investigation of the critical region of CNVs often leads to discovery of novel ID-associated genes such as *MBD5* and *KANSL1* (156, 157). There is also evidence that some cases of ID are caused by a ‘two-hit’ model, where two different CNVs are required for manifestation of disease (158). This finding blurs the dichotomy between monogenic and polygenic models of disease.

Historically, discovery of single gene causes of ID was largely limited to families with a typical pattern of X-linked inheritance. *FMR1* was the first X-linked ID-associated gene to be identified, by positional mapping of yeast artificial chromosome clones followed by Sanger sequencing (159). Triplet expansion repeats within *FMR1* cause fragile X syndrome, which is the most common single gene cause of ID, accounting for ~0.5% of cases (160). Another important example of an X-linked ID-associated gene is *MECP2*, pathogenic variants in which were originally found to be the cause of Rett syndrome in females. Pathogenic *MECP2* variants have since been implicated in a variety of forms of ID in both males and females, although they are a much more common cause in females (161). Single gene, X-linked ID accounts for ~10% of ID cases overall (162). A study published in 2009 illustrated the importance of large-scale sequencing in the discovery of ID-associated genes (13). The authors recruited 208 families with X-linked ID, and sequenced 65% of all the coding regions of the X chromosome by Sanger sequencing. This was the largest systematic screen for pathogenic variants at the time, and discovered nine novel X-linked ID-associated genes including *CASK*.

The widespread availability of next generation sequencing (NGS) that flourished very shortly after the publication of the study just described, opened up possibilities of ID-associated gene discovery on a whole new scale. For the first time, autosomal single nucleotide variants (SNVs) and insertion deletions (indels) that cause ID could be identified systematically. To achieve this, one study performed exome sequencing in 136 consanguineous families affected by autosomal recessive ID (163). Homozygosity mapping in each family allowed the analysis to be restricted to loci likely to contain the causative variant. As well as identifying pathogenic variants in known ID-associated genes, 50 possible novel ID-associated genes were identified, including some that have subsequently been confirmed, including *KIF7*, *MAN1B1*, and *TAF2*.

Exome sequencing using a trio study design has repeatedly shown that *de novo* mutations are a very important cause of ID, and account for a large proportion of cases (57, 84, 164). The *de novo* paradigm of ID along with the reduced reproductive fitness of ID patients probably explains the long known observation that many forms of ID occur sporadically. Also, it explains the apparent paradox between the relatively high prevalence of ID, and the fact that it significantly reduces reproductive fitness.

Whole genome sequencing can identify coding pathogenic variants that were missed by exome sequencing (62). Admittedly, the exome sequencing in the original study may have called variants with lower sensitivity than subsequent studies, as demonstrated by the low *de novo* exome mutation rate of 0.53 per patient (57). Nevertheless, whole genome sequencing has fewer biases in variant calling, and greater uniformity of coverage than whole exome sequencing, suggesting that an eventual move away from exome sequencing towards whole genome sequencing is likely (63).

3.1.3 Biology of intellectual disability-associated genes

Over 500 single genes in which pathogenic variants may cause ID have been identified thus far, with many more unconfirmed candidates (62). ID is so genetically heterogeneous that it is appropriate to consider the term to be a hypernym describing many individual syndromes and non-syndromic forms (165). ID-associated genes may be classified and understood according to their function, or the pathway in which they act (Table 3-1) (135). This is helpful because it facilitates identification of further candidate genes, and helps with prognosis, and because pathogenic variants in different genes in the same pathway often cause similar phenotypes. Some functional classes affect universal cellular processes, whereas others are very specific to neurological processes.

Functional class of ID-associated gene	Examples	References
Presynaptic vesicle release and recycling	*STXBP1; CASK; IL1RAPL1*	(166-168)
Neurotransmitter receptors	*GRIA3; GRIN2A; GRIN2B*	(101, 169)
Components of the post-synaptic density	*SYNGAP1; SHANK2*	(170, 171)
Regulators of gene expression	*MECP2; EHMT1; ARID1B; FMR1*	(161, 172-174)
Metabolism	*PAH; PMM2*	(175, 176)

Table 3 - 1: Functional classes of ID-associated genes.

At a typical synapse, the presynaptic terminal contains vesicles filled with neurotransmitter. The primary excitatory neurotransmitter is glutamate, and the primary inhibitory neurotransmitter is γ-aminobutyric acid (GABA). When stimulated, the vesicles fuse with the presynaptic membrane and exocytose their contents into the synaptic cleft, whereupon the cell recycles the vesicles. The release and recycling of pre-synaptic vesicles are complex biological processes involving many proteins. Pathogenic variants in genes that encode some of these proteins can cause ID. For example, *de novo* mutations in *STXBP1* can cause Ohtahara syndrome (166). *STXBP1* encodes Munc18-1, a protein required for fusion of the vesicles with the presynaptic membrane. CASK is also involved in exocytosis (167). IL1RAPL1, on the other hand, *inhibits* neurotransmitter release; pathogenic variants in *IL1RAPL1* can cause non-syndromic X-linked ID, ASD or schizophrenia (168).

In the synaptic cleft, neurotransmitters bind to receptors on the postsynaptic membrane on dendritic spines of the neuron receiving the signal. For excitatory synapses, the two main types of glutamate receptors are NMDA and AMPA receptor. Pathogenic variants in genes that encode subunits of these receptors can cause ID. For example, variants in *GRIA3*, which encodes a subunit of the AMPA receptor, can cause moderate X-linked ID (169). Similarly, *de novo* mutations in *GRIN2A* or *GRIN2B*, which encode subunits of the NMDA receptor, can cause ID and seizures (101).

Neurotransmitter receptors are part of an extensive protein complex called the postsynaptic density (PSD). Proteins in this complex perform many functions from regulating and propagating the signal, to providing structural support to the receptors. Integrity of the PSD is required for various cognitive processes including learning and memory. It is therefore unsurprising that mutations in PSD proteins other than the receptors themselves (such as the regulatory protein SYNGAP1 or the scaffolding protein SHANK2) can cause ID and other neurodevelopmental disorders (170, 171).

People with ID may or may not have structural brain abnormalities apparent on imaging such as magnetic resonance imaging (MRI). Regardless of this, histology on post-mortem brain samples often shows characteristic changes to the structure of dendrites and dendritic spines compared to healthy people, although it is unclear whether this is a cause or a consequence of the cognitive defect (177). Plasticity of dendritic spine morphology is important for cognitive functioning. Rapid changes to dendritic spine morphology are achieved by remodelling of actin filaments and microtubules. Pathogenic variants in genes that encode proteins that regulate this remodelling process can therefore cause ID (including *OPHN1* and *FGD1*) (178, 179).

Glutamate binding to NMDA or AMPA receptors activates signaling cascades such as the RAS-MAPK pathway in the postsynaptic neuron. Pathogenic variants in members of this pathway cause a family of diseases that are becoming known as RASopathies, one common feature of which is ID. For example, *de novo* mutations in *HRAS* can cause Costello syndrome (180). Typical features of Costello syndrome are ID, short stature, excess skin and dysmorphic craniofacial features. Interestingly, RASopathies may potentially be one class of ID that could benefit from pharmaceutical intervention (181).

Another important class of ID-associated genes is regulators of gene expression. Appropriate transcription and translation of downstream genes is necessary for cognitive function. This is demonstrated by the fact that pharmaceutical inhibition of protein synthesis using an agent such as anisomycin inhibits the formation of memories (182). Several functional classes of genes regulate gene expression. Transcription factors do so by directly binding to DNA response elements, histone modifiers catalyse the addition or removal of groups (e.g. acetyl or methyl groups) to or from histone proteins, and DNA methyltransferases catalyse the transfer of methyl groups onto DNA itself. Transcription regulators are increasingly recognised as an important cause of ID. The problem with understanding how they do so is that usually the downstream genes whose expression is altered are not known. MECP2 is a transcription regulator that binds to the methylated DNA response element of a downstream gene and initiates formation of a complex that silences the gene. Euchromatic histone methyltransferase 1, encoded by *EHMT1*, catalyses the transfer of methyl groups onto lysine residues of histone proteins and is particularly enriched in brown adipose tissue (183). Disruption of *EHMT1* can cause Kleefstra syndrome, where patients have ID, hypotonia, brachycephaly, dysmorphic facial features, and CHD (172). Similarly, heterozygous *de novo* mutations in the SWI/SNF chromatin remodelling complex component *ARID1B* are a more frequent cause of ID, accounting for ~1% of previously undiagnosed cases (173). FMRP, which is encoded by *FMR1*, is an RNA-binding protein that regulates the expression of other proteins including components of the PSD (174).

Finally, pathogenic variants in metabolic genes can cause inborn errors of metabolism (IEM), a common feature of which is ID. For example, recessive variants in *PAH*, which encodes phenylalanine hydroxylase, cause phenylketonuria (PKU). Patients with untreated PKU have ID, seizures, microcephaly and hypopigmentation (175). Most developed countries have implemented screening programs for PKU, and treat patients from infancy with dietary changes and medication. Another example of an IEM where ID is a feature is congenital disorder of glycosylation (CDG). Here, pathogenic recessive variants in genes such as *PMM2*, which are involved in glycosylation of downstream proteins, cause phenotypes including ID, cardiomyopathy, frequent infections, central nervous system and eye defects (176).

Interestingly, the functional class of a gene can affect aspects of the associated disease, such as the mode of inheritance. Genes associated with IEMs have recessive inheritance, whereas genes encoding chromatin modifiers are usually haploinsufficient, so pathogenic variants cause disease with dominant inheritance (172, 173, 175, 176). Intuitively, it seems likely that pathogenic variants in ID-associated genes that are very specific to neurological processes might, on average, cause ID that is largely non-syndromic, whereas pathogenic variants in ID-associated genes that affect universal cellular processes might cause a more syndromic phenotype, because more systems will be affected. While some of the examples I have given in this section support this hypothesis, others do not. More large-scale, unbiased studies of ID are required to establish whether this pattern exists.

The current diagnostic yield for ID patients is up to 50-65% (62, 135, 152). There are several reasons why 35-50% of ID patients still do not receive a genetic diagnosis, including the possibility that the causative variant could be in a non-coding region, it could be in a gene not known to be ID-associated, or the disease could be caused by several variants acting in an oligogenic manner. Nevertheless, it is clear that more ID-associated genes remain to be identified.

3.1.4 Case-control enrichment analysis of rare variants

Rare disease-associated genes are usually identified by means of a classical, case-only diagnostic approach, where they are identified because they contain rare, coding variants which segregate with disease in multiple families, for example. Case-control enrichment analysis is a supplementary method that can yield additional insights into the aetiology of rare disease. Typically, a cohort of cases is assembled, along with a cohort of controls. Rare variants are identified in both cohorts (for example by exome sequencing), and then a statistical test is applied to test the hypothesis that the cases have an excess of a defined category of variants compared to controls. Case-control enrichment analysis can yield insights into the genetic architecture of a rare disease without necessarily assigning causality to individual variants. It can be used with a range of study designs, whereas classical approaches often require very specific study designs. For example, to identify *de novo* mutations DNA samples from both biological parents are required, which are not always available. Perhaps most importantly, case-control enrichment analysis makes fewer assumptions about causative variants than classical approaches, and therefore takes into account non-classical contributors to disease such as variants with incomplete penetrance, and variants that contribute to a phenotype in an oligogenic manner.

Several different statistical tests have been developed for use in case-control enrichment analysis (recently reviewed in (184)). Three of the most commonly used are the cohort allelic sums test (CAST), the weighted sum method, and the sequence kernel association test (SKAT). CAST is a burden test, whereby information about a variant category of interest is collapsed into whether each individual has any variant of that category, or whether they do not. A statistical test (usually Fisher's exact test) is then applied to this count data to assess the degree and significance of any difference between the cohorts. CAST was first formally described in 2007 (185), although it had been used prior to this (186). It is a very flexible test, in that it can be used to test for association between an individual gene and a phenotype, association between a group of genes and a phenotype, or even a genome wide burden of variants. A disadvantage of the CAST test is that it assumes that the direction and size of effect of all variants are the same. If this assumption is not true, power is lost. CAST also assumes that a fairly large proportion of variants are causal. Also, by collapsing information, power is lost. For example, an individual with ten rare variants of interest is treated with the same weight as an individual with only one such variant, whereas it may be more appropriate for the individual with ten variants to be given a higher weight in the test. However, where the assumptions are true, CAST is a robust and powerful test (185).

Because of the assumption of CAST that a relatively high proportion of variants are pathogenic, prior to performing CAST, filtering based on minor allele frequency should be performed (a typical cutoff is 0.01). However, then a unique variant is still treated with the same weight as a variant with a frequency of 0.01, whereas it may be more appropriate for the unique variant to be given a higher weight in the test. The weighted sum method is very similar to CAST, but variants of all frequencies are included, and collapsed into a single average number of rare alleles per case, weighted according to variant frequency in controls (187). Therefore, the weighted sum method has greater power than CAST if one wants to simultaneously test variants of different frequencies. The weighted sum method makes the same assumptions as CAST about direction and size of effect.

SKAT is a variance-component test, which uses a regression framework to evaluate differences in the distribution of various scores between variants in cases and controls without collapsing the information into a single statistic (188). It is flexible, computationally efficient, can account for covariates, and makes no assumptions about direction and size of genetic effect. Where a phenotype is influenced by variants with different directions of effect, SKAT is much more powerful than CAST or the weighted sum method. However, where the effects are in the same direction, and most variants are pathogenic, CAST is more powerful (184).

Regardless of the statistical test selected, case-control association or enrichment analyses are potentially subject to spurious findings if there are systematic differences between cases and controls. These can be technical differences, if, for example, the cases and controls were sequenced in different batches. Population stratification between cases and controls can lead to differences in allele frequency that can falsely appear as a disease association (189). One commonly used method by which to detect, and if necessary adjust for, population stratification is principal component analysis (PCA).

Several studies demonstrate the utility of case-control enrichment analysis in understanding the role of variants in rare disease. In an early example, Cohen *et al.* Sanger sequenced the coding regions of three genes in which pathogenic variants can cause Mendelian forms of low high-density lipoprotein cholesterol (HDL-C), in individuals from the general population with low HDL-C levels, compared to controls with high HDL-C levels (186). They used the CAST test to demonstrate that individuals with low HDL-C levels had a significant burden of rare non-synonymous variants in the candidate genes compared to the high HDL-C controls, suggesting some shared aetiology between Mendelian forms of low HDL-C, and low HDL-C in the general population. In another important example, Cooper *et al.* identified an enrichment of rare, large (>400 kb) CNVs in children with ID compared to controls (155).

In 2013, Liu *et al.* whole exome sequenced a cohort of over 1000 ASD patients, along with 870 controls (190). The authors used both the weighted sum test and the SKAT test in an attempt to identify novel ASD-associated genes, with an excess of rare, coding variants in cases. They did not find any genes, known or novel, with an exome-wide significant burden, demonstrating that much larger sample sizes are required for gene discovery using this method.

Purcell *et al.* recently took a slightly different approach, in order to investigate the genetic aetiology of schizophrenia (14). Instead of looking for a burden in individual genes, the authors took a ‘top-down’ approach, and focused on groups and subgroups of candidate genes. This method increased their power to detect an enrichment of variants, and simultaneously reduced the burden of multiple testing, which proved to be successful. Using a combination of the CAST and SKAT tests on exome sequencing data, the authors identified an enrichment of rare coding variants in candidate schizophrenia genes in patients with schizophrenia compared to controls. A particularly large enrichment was identified in components of the postsynaptic activity-regulated cytoskeleton-associated scaffold complex, emphasising the importance of this complex in the aetiology of schizophrenia.

3.1.5 Aims, context, and colleagues

The overall aims of this project were threefold. The first aim was to identify pathogenic loss of function (LOF) and missense variants in known ID-associated genes in ID patients, the second was to identify novel ID-associated genes, and the third was to determine whether there is a significant enrichment of variants in ID-associated genes in ID patients compared to controls. These aims were addressed by means of a targeted resequencing study of rare diseases that was carried out as part of the UK10K project. This project was a large collaborative effort. In this chapter, I have included a few instances of work done by other people, where it is necessary to put my own work into context. I have made it clear who did the work at the point I describe it, and I also summarise it here.

The UK10K rare disease consortium, chaired by Dr Matthew Hurles and Dr David Fitzpatrick designed and implemented the study. Dr Lucy Raymond led the ID cohort, and along with Dr Detelina Grozeva and Dr Olivera Spasic-Boskovic assembled and prepared samples, selected ID-associated genes to be sequenced, did the case-only diagnostic analysis, novel gene identification, and validations. An international collaborative team of clinicians and researchers including Dr Michael Parker, Dr Hayley Archer, Dr Helen Firth, Dr Soo-Mi Park, Dr Natalie Canham, Dr Susan Holder, Dr Meredith Wilson, Dr Anna Hackett, and Dr Michael Field contributed samples to the ID cohort. Professor Shoumo Bhattacharya, Dr Jamie Bentham, and Dr Catherine Cosgrove assembled the CHD cohort. Dr James Floyd designed the custom sequencing pull-down experiment and performed quality control analysis on the data. The high-throughput sequencing team at the Wellcome Trust Sanger Institute (WTSI) did the DNA amplification, pull-down and sequencing. Dr Shane McCarthy led the initial bioinformatics including read mapping and variant calling. Dr Saeed Al Turki wrote some Python scripts that I used during this project.

The parts of the project for which I was responsible are as follows: annotating variants, designing and implementing a filtering pipeline to identify possibly causative variants, assisting with interpretation of data to identify causative variants and novel genes, and designing and performing an extensive series of burden tests to investigate the extent to which variants in ID-associated genes are enriched in ID patients (including PCA and CAST). I carried out this work under the supervision and guidance of Dr Matthew Hurles.

Some parts of this chapter have been published ((191) and manuscript in preparation). Unless otherwise stated, where material in this chapter is taken from those publications, I declare that those sections were originally my own work.

3.2 Methods

3.2.1 Samples, sequencing, and quality control

Genomic DNA of 2812 individuals with one of seven rare diseases was whole genome amplified using 1µl of 10ng/µl template DNA using GenomiPhi kit (GE Healthcare). Dr James Floyd designed custom targeted Agilent SureSelect pull-down baits using the SureDesign program (Agilent Technologies, Santa Clara, CA, USA). This targets 3.35 Mb of sequence from the coding exons (GRCh37) of 1189 genes. These genes consist of candidates for each of the seven rare diseases. The 565 sequenced ID-associated genes were selected by Dr Lucy Raymond, and an international collaborative team of clinical geneticists assembled the ID samples. Target enrichment was done using a custom SureSelect library (Agilent Technologies), according to the manufacturer's instructions. The Illumina HiSeq 2000 platform (Illumina, Inc. San Diego, USA) was used to perform the sequencing. The high-throughput sequencing team at WTSI did the DNA amplification, pull-down and sequencing.

Dr Shane McCarthy at WTSI led the work described in this paragraph. Each read was aligned to the reference genome (GRCh37) using the Burrows-Wheeler Alignment tool, and SNVs and indels were identified using both SAMtools mpileup and the GATK UnifiedGenotyper, and these variants calls were merged, prioritising GATK calls at sites where there was a discrepancy (76, 192). Variants were stored in variant call format (VCF) files both as single-sample and multi-sample calls. Functional annotations were added using the Ensembl Variant Effect Predictor v2.8 against Ensembl 70 (193). Additionally, some basic filters were applied to the variants; including removal of very low coverage calls, and calls where the reference base is unknown. Dr James Floyd generated and analysed quality control metrics.

3.2.2 Annotation and filtering pipeline

I used a python script written by Dr Saeed Al Turki to add minor allele frequency data to each variant from the following sources: 1000 genomes database, UK10K twins cohort, exome sequencing project (ESP) 6500, and a cohort of 2172 control individuals exome sequenced at WTSI. I wrote an R script to calculate and annotate the internal variant frequency (the frequency with which each variant appeared in the UK10K replication study, including all phenotypes).

I designed and implemented a filtering pipeline using R, to generate a list of rare, possibly causative variants from the merged and annotated VCF files. I only considered variants that had minor allele frequency < 0.01 in all four databases described, internal frequency < 0.01, quality score > 40, and mapping quality score > 50. I selected the quality score and mapping quality score cutoffs by visually inspecting the original sequencing data of a subset of variants using The Integrative Genomics Viewer (IGV) (82). I also removed heterozygous calls on the X chromosome in males. I selected the most severe consequence of each variant, and considered only variants with two categories of consequence: functional (coding sequence variants, in-frame deletions, in-frame insertions, initiator codon variants, missense variants, and variants resulting in loss of a stop codon) or predicted LOF (nonsense, frameshift or essential splice site variants). Finally, I considered only variants in the sequenced ID-associated genes.

To determine whether there was an excess of *de novo* LOF mutations in a particular gene, I calculated the number expected to occur by chance using the known exome mutation rate, the proportion of mutations that are expected to be LOF, and taking into account the length of the coding sequence of the gene (83, 84). I compared this to the observed number of *de novo* LOF mutations, assuming a Poisson distribution to calculate a p-value, which I corrected for testing of multiple genes using the Bonferroni correction.

3.2.3 Principal component analysis

PCA was done using the R package SNPRelate, which is a convenient and computationally efficient tool (194). I converted VCF files of multi-sample calls for each of the ID and CHD cohorts to GDS format using the snpgdsVCF2GDS function of the SNPRelate package. I used the snpgdsLDpruning function of the SNPRelate package to identify a list of 2291 high-quality, biallelic and polymorphic SNVs with minor allele frequency ≥0.05, that are not in linkage disequilibrium (LD) with each other, in the UK10K samples. Next, I performed PCA on the UK10K samples along with a subset of unrelated HapMap3.3 samples, using the snpgdsPCA function of the SNPrelate package and the 2291 SNVs identified (195).

3.2.4 Cohort allelic sums test

I wrote an R script that reads in a file of variants in sequenced ID-associated genes in the ID cohort and CHD controls. The script identifies and removes samples with excessive numbers of variants. Additional filters can then be applied to the variants, for example to remove those which have an IGV plot suggestive of a false positive call, or to apply more stringent internal frequency cutoffs. Next, the variants are subcategorised. I classified the 565 genes into known (n=204; Table 3-2) and candidate (n=361; Table 3-3) according to whether they are present in a stringent, manually curated list of known ID-associated genes in a recently published study (62). The script counts the number of variants in each sample, and generates a 2x2 contingency table where each row is one of the two cohorts, and the two columns respectively show the number of samples who have at least one variant, and the number who do not. Finally, a one-tailed Fisher's exact test is performed on the contingency tables.

ABCD1	*ADCK3*	*ADSL*	*AFF2*	*AGA*	*AGTR2*
ALDH18A1	*ALDH5A1*	*ALG1*	*ALG12*	*ALG3*	*ALG6*
ANK3	*AP1S2*	*ARFGEF2*	*ARHGEF9*	*ARID1A*	*ARID1B*
ARX	*ASXL1*	*ATP7A*	*ATRX*	*AUH*	*BCOR*
BRAF	*CASK*	*CC2D2A*	*CCDC22*	*CDH15*	*CDKL5*
CEP41	*CHD2*	*CHD7*	*CNTNAP2*	*CREBBP*	*CTNNB1*
CUL4B	*DCX*	*DHCR7*	*DKC1*	*DLG3*	*DMD*
DNMT3B	*DYNC1H1*	*DYRK1A*	*EHMT1*	*EP300*	*ERCC6*
EXOSC3	*FGD1*	*FKRP*	*FKTN*	*FLNA*	*FMR1*
FOXG1	*FOXP1*	*FTSJ1*	*GCH1*	*GDI1*	*GJC2*
GK	*GPC3*	*GPR56*	*GRIA3*	*GRIN2A*	*GRIN2B*
HCCS	*HCFC1*	*HDAC4*	*HDAC8*	*HPRT1*	*HRAS*
HSD17B10	*HUWE1*	*IDS*	*IDUA*	*IKBKG*	*IL1RAPL1*
INPP5E	*IQSEC2*	*KANK1*	*KANSL1*	*KAT6B*	*KCNQ3*
KDM5C	*KIF7*	*KIRREL3*	*KRAS*	*L1CAM*	*LAMP2*
LRP1	*LRP2*	*MAP2K1*	*MAP2K2*	*MBD5*	*MECP2*
MED12	*MEF2C*	*MID1*	*MLH1*	*MLL2*	*MLL3*
MLYCD	*MMAA*	*MMADHC*	*MYT1L*	*NDE1*	*NDP*
NEU1	*NF1*	*NFIX*	*NHS*	*NLGN4X*	*NRXN1*
NSD1	*NSDHL*	*NSUN2*	*OCRL*	*OFD1*	*OPHN1*
OTC	*PAFAH1B1*	*PAK3*	*PARP1*	*PAX6*	*PC*
PCDH19	*PCNT*	*PDHA1*	*PEPD*	*PGK1*	*PHF6*
PHF8	*PLP1*	*PNKP*	*POLR3A*	*POLR3B*	*PORCN*
PRPS1	*PTCHD1*	*PTEN*	*PTPN11*	*RAB3GAP1*	*RAF1*
RAI1	*RPS6KA3*	*SATB2*	*SCN2A*	*SCN8A*	*SETBP1*
SETD5	*SHANK2*	*SHANK3*	*SHOC2*	*SHOX*	*SHROOM4*
SLC12A6	*SLC16A2*	*SLC26A9*	*SLC2A1*	*SLC6A8*	*SLC9A6*
SMARCA2	*SMARCA4*	*SMARCB1*	*SMARCE1*	*SMC1A*	*SMS*
SOS1	*SOX3*	*SOX5*	*SPRED1*	*SPTAN1*	*SRGAP3*
STXBP1	*SYN1*	*SYNE1*	*SYNGAP1*	*SYP*	*TAT*
TBC1D24	*TCF4*	*TIMM8A*	*TRAPPC9*	*TSC1*	*TSC2*
TSPAN7	*TUBA1A*	*TUBB2B*	*TUSC3*	*UBE2A*	*UBE3A*
UBR1	*UPF3B*	*VLDLR*	*VPS13B*	*WDR11*	*WDR62*
ZDHHC9	*ZEB2*	*ZFHX4*	*ZFYVE26*	*ZNF41*	*ZNF674*

Table 3 - 2: List of 204 sequenced intellectual disability-associated genes that are known. Genes were classified as known if they are in a stringent, manually curated list of known ID-associated genes from a recently published study (62). *SETD5* was included in this list for the purposes of the case-control enrichment analyses, as a result of findings described in this chapter.

ACBD6	*ACE2*	*ACIN1*	*ACOT9*	*ACSL4*	*ACTL6A*
ACTL6B	*ACY1*	*ADK*	*ADRA2B*	*AIMP1*	*AKAP17A*
AKAP4	*ALDH4A1*	*ALG13*	*ALG8*	*AP4B1*	*AP4E1*
AP4M1	*AP4S1*	*ARG1*	*ARHGAP36*	*ARHGAP6*	*ARHGEF4*
ARHGEF6	*ARID2*	*ARIH1*	*ARL14EP*	*ARSF*	*ASB12*
ASCC3	*ASCL1*	*ASH1L*	*ASMT*	*ASMTL*	*ATM*
ATP2B3	*ATXN3L*	*AVPR2*	*AWAT2*	*BCORL1*	*BDP1*
BMP15	*BRWD3*	*BTK*	*C12orf57*	*CA8*	*CACNA1F*
CACNA1G	*CAMK2A*	*CAMK2G*	*CAP1*	*CAPN10*	*CASP2*
CC2D1A	*CCDC23*	*CCNA2*	*CCNB3*	*CD99*	*CDK16*
CDK8	*CFP*	*CHL1*	*CLCN4*	*CLCN5*	*CLIC2*
CMC4	*CNKSR1*	*CNKSR2*	*COL4A3BP*	*COL4A6*	*COQ5*
COX10	*CPXCR1*	*CRLF2*	*CSF2RA*	*CSTF2*	*CTPS2*
CTSD	*CTTNBP2*	*CUX2*	*CXORF22*	*CXORF58*	*CYP7B1*
DCHS2	*DDOST*	*DDX26B*	*DDX3X*	*DDX53*	*DEAF1*
DGKH	*DGKK*	*DHRSX*	*DHX30*	*DIAPH2*	*DLG1*
DLG2	*DLG4*	*DOCK11*	*DPF1*	*DPF2*	*DPF3*
EEF1A2	*EEF1B2*	*EIF2C1*	*EIF2S3*	*ELK1*	*ELP2*
ENOX2	*ENTHD2*	*ENTPD1*	*EPPK1*	*ERLIN2*	*ESX1*
FAAH2	*FAM120C*	*FAM47B*	*FAM58A*	*FASN*	*FKBPL*
FRMPD4	*FRY*	*FTL*	*GAB3*	*GABRQ*	*GAD1*
GATAD2B	*GCDH*	*GLB1*	*GLRA2*	*GM2A*	*GON4L*
GPR112	*GPRASP1*	*GRB14*	*GRIA1*	*GRIA2*	*GRIK2*
GSPT2	*GTPBP8*	*HAUS7*	*HDHD1*	*HEXA*	*HEXB*
HGSNAT	*HIST1H4B*	*HIST3H3*	*HIVEP2*	*HS6ST2*	*HSPD1*
IFNAR2	*IGSF1*	*IL3RA*	*INPP4A*	*ITGA4*	*ITIH6*
KCNC3	*KCND1*	*KCNH1*	*KCNK12*	*KDM1A*	*KDM5A*
KDM6B	*KIAA2022*	*KIF1A*	*KIF26B*	*KIF4A*	*KIF5C*
KLHL15	*KLHL21*	*KLHL34*	*KLHL4*	*LAMA1*	*LARP7*
LAS1L	*LHFPL3*	*LIMK1*	*LINS*	*LRRK1*	*MAGEA11*
MAGEB1	*MAGEB10*	*MAGEB2*	*MAGEC1*	*MAGEC3*	*MAGED1*
MAGEE2	*MAGIX*	*MAGT1*	*MAN1B1*	*MAOA*	*MAOB*
MAP3K15	*MAP7D3*	*MBNL3*	*MED17*	*MED23*	*MGAT5B*
MIB1	*MLC1*	*MMAB*	*MORC4*	*MSL3*	*MTF1*
MTMR1	*MTMR8*	*MXRA5*	*MYO1D*	*MYO1G*	*NAA10*
NDST1	*NDUFA1*	*NECAB2*	*NKAP*	*NLGN3*	*NR1I3*
NRK	*NRXN2*	*NTM*	*NXF4*	*NXF5*	*ODF2L*
OGT	*OR5M1*	*OXCT1*	*P2RY4*	*P2RY8*	*PABPC5*
PAH	*PASD1*	*PBRM1*	*PCDH10*	*PECR*	*PGRMC1*
PHACTR1	*PHF10*	*PHIP*	*PHKA1*	*PIGN*	*PIK3C3*
PIN4	*PJA1*	*PLA2G6*	*PLCXD1*	*PLXNB3*	*POLA1*
PPP2R5D	*PPT1*	*PQBP1*	*PRDX4*	*PRICKLE3*	*PRMT10*
PROX2	*PRRG1*	*PRRG3*	*PRRT2*	*PRSS12*	*PSMA7*
PSMD10	*PTPN21*	*RAB39B*	*RAB40AL*	*RABL6*	*RALGDS*
RAPGEF1	*RBM10*	*RENBP*	*RGAG1*	*RGN*	*RGS7*
RLIM	*RNASET2*	*RPGR*	*SCAPER*	*SETDB2*	*SGSH*
SHANK1	*SHROOM2*	*SLC25A22*	*SLC25A53*	*SLC25A6*	*SLC31A1*
SLC6A1	*SLC6A17*	*SMARCC1*	*SMARCC2*	*SMARCD1*	*SMARCD2*
SMARCD3	*SNTG1*	*SPG11*	*SPRY3*	*SPTLC2*	*SREBF2*
SRPX2	*ST3GAL3*	*STAB2*	*STAG1*	*STARD8*	*SYNCRIP*
SYT1	*SYTL4*	*SYTL5*	*TAF1*	*TAF2*	*TAF7L*
TANC2	*TBC1D8B*	*TCEAL3*	*TCP10L2*	*TENM1*	*THAP1*
THOC2	*ThumpD1*	*TKTL1*	*TLR8*	*TM4SF2*	*TMEM132E*
TMEM135	*TMLHE*	*TNKS2*	*TNPO2*	*TREX2*	*TRIO*
TRMT1	*TSC22D3*	*TSEN2*	*TSEN34*	*TSEN54*	*TTI2*
TUBA8	*TUBAL3*	*UBR7*	*UBTF*	*USP27X*	*USP9X*
UTP14A	*VAMP7*	*VRK1*	*WAC*	*WDR13*	*WDR45L*
WNK3	*WWC3*	*XIAP*	*XKRX*	*YY1*	*ZBTB40*
ZC3H14	*ZCCHC12*	*ZCCHC8*	*ZDHHC15*	*ZFX*	*ZMYM3*
ZMYM6	*ZMYND12*	*ZNF238*	*ZNF425*	*ZNF526*	*ZNF711*
ZNF81					

Table 3 - 3: List of 361 sequenced intellectual disability-associated genes that are candidates.

Genes were allocated as candidate if they are not in a stringent, manually curated list of known ID-associated genes from a recently published study (62).

3.3 Results

3.3.1 Targeted resequencing of 565 intellectual disability-associated genes in cases and controls

The coding regions of a set of 565 known or candidate ID-associated genes were sequenced in 996 individuals (94% male) with moderate to severe, sporadic ID. This was a subset of a large replication study of seven rare diseases, comprising a total of 2812 individuals, which was carried out within the UK10K study (www.UK10K.org). The phenotypes studied were CHD, ciliopathy, coloboma, ID, neuromuscular disease, severe insulin resistance, and congenital thyroid disease, along with internal technical control samples. Coding regions of a total of 1189 genes (of which 565 are known or candidate ID-associated genes) were selected using a custom pull-down approach, then sequenced on the Illumina HiSeq 2000 platform.

The 565 ID-associated genes included known genes in which pathogenic variants in multiple unrelated individuals have been shown to cause ID, and also candidate genes selected, for example, because a variant has been identified in a single patient with ID, or because the gene is in the same family as known ID-associated genes. Some recently published studies of ID have larger lists of ID-associated genes (62). This is because some new ID-associated genes have been identified since the design of our study, and also because of restrictions on the size of targeted regions imposed by the pull-down method. I classified the 565 genes into known and candidate genes.

3.3.2 The sequencing data are of good quality

There are around 1500 coding SNVs and 50 coding indels per sample in this study that pass standard quality control filters (Figure 3-1). The mean depth of variant coverage per sample is 40.55X (Figure 3-1c). This is higher than the minimum 30X estimated to be required for accurate detection of heterozygous variants (87). Dr James Floyd calculated these figures.

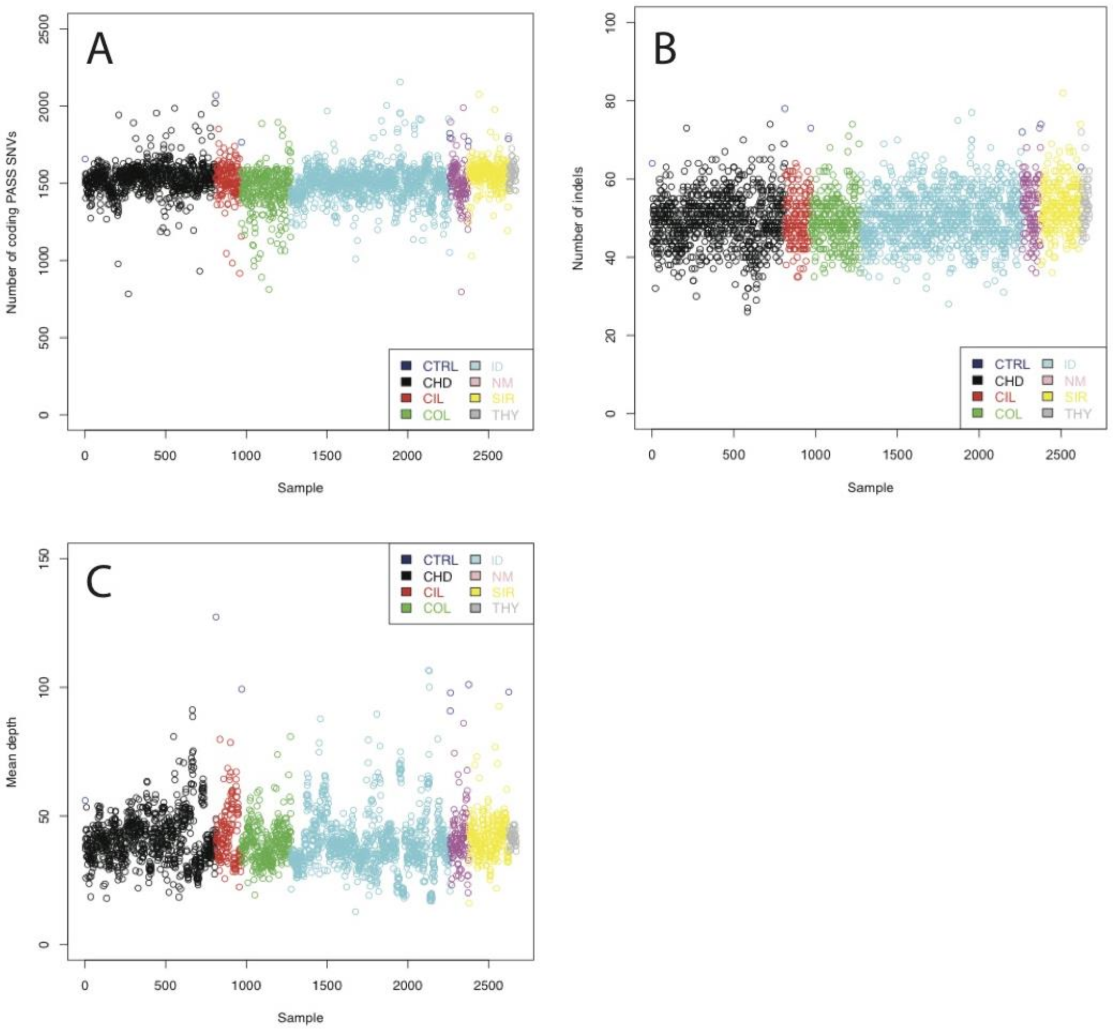

Figure 3 - 1: Quality control metrics for the UK10K targeted resequencing study.
A) Number of pass, coding SNVs per sample, across 1189 sequenced genes. **B)** Number of coding indels per sample, across 1189 sequenced genes. **C)** Mean depth of variant coverage for each replication sample. CTRL = controls, CHD = congenital heart disease, CIL = ciliopathy, COL = coloboma, ID = intellectual disability, NM = neuromuscular disorders, SIR = severe insulin resistance, THY = thyroid disease. Numbers and plots generated by Dr James Floyd, and included here with permission.

3.3.3 There is no substantial difference in population structure between the intellectual disability and congenital heart disease cohorts

Identification of an enrichment of predicted damaging variants in selected disease-associated genes in individuals with the disease of interest, compared to controls, often leads to insights about the disease pathology (14). I hypothesised that there might be an excess of variants in sequenced ID-associated genes in the ID part of the UK10K rare disease cohort, compared to controls. For controls, I selected the CHD cohort. This is an appropriate control because the two cohorts are of similar size, and have minimal overlap in phenotypic spectra. The CHD DNA samples had been treated, stored, amplified, sequenced, and analysed in an identical manner to those of the ID cohort.

However, population stratification between cohorts can lead to spurious findings in case-control analyses (189). The majority of the ID and CHD cohorts reported as being of European ancestry. Nevertheless, to find out whether there was a substantial difference in population structure between the two cohorts I used PCA. This is a widely used method for this purpose (184). I performed PCA on the ID and CHD samples, along with a subset of unrelated HapMap3.3 samples, using the SNPrelate package (195).

The first two principal components were sufficient to cluster the HapMap samples into their four component populations (Figure 3-2). The data points for the UK10K ID cases and CHD controls overlie each other, suggesting that there is no substantial difference in population structure between these cohorts. Additionally, they overlap to a large extent with the data points from the HapMap samples of European ancestry, confirming that the majority of both the ID cases and CHD controls are of European ancestry.

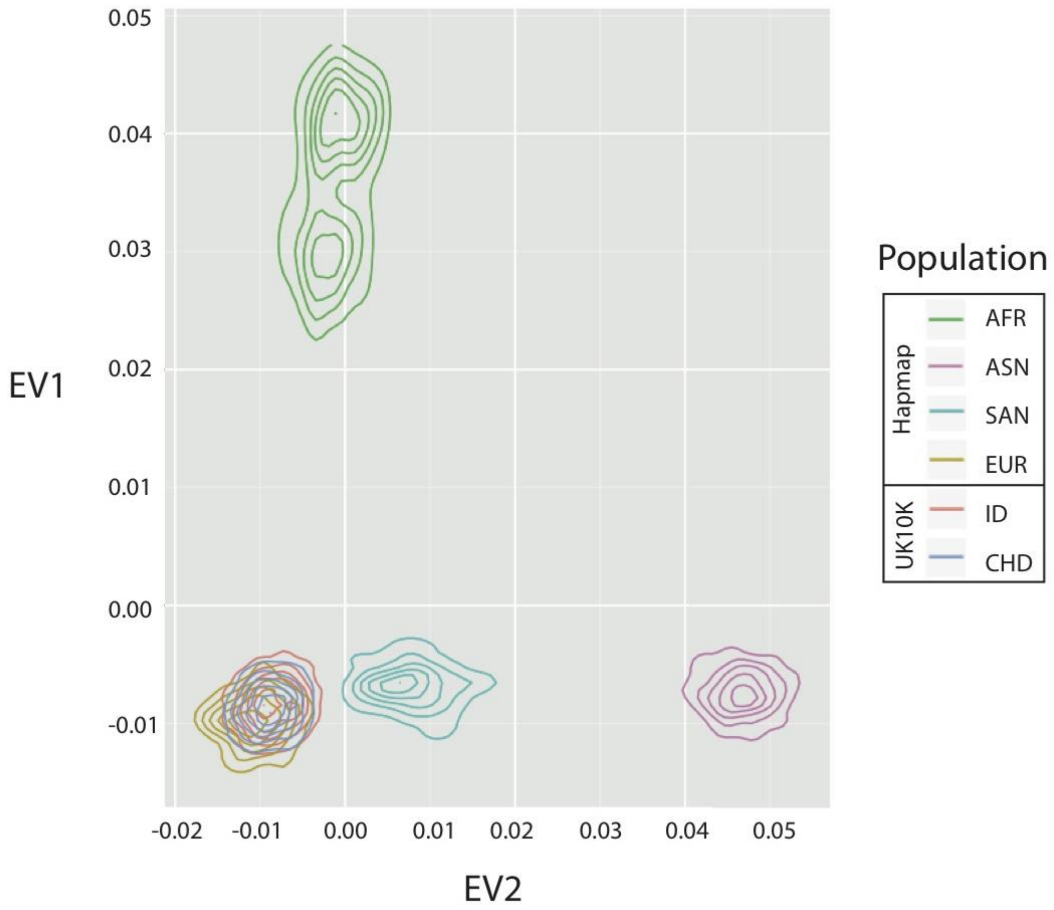

Figure 3 - 2: Principal component analysis.
The first two eigenvectors (EVs) cluster the HapMap3.3 samples into their component populations (AFR = individuals of African ancestry; ASN = individuals of East Asian ancestry; SAN = individuals of South Asian ancestry; EUR = individuals of European ancestry) (195). The UK10K ID and CHD samples overlie with each other, and overlap with the European HapMap3.3 samples.

3.3.4 14% of intellectual disability patients have a likely causative variant in a sequenced intellectual disability-associated gene

I wrote a set of R scripts to generate a list of rare, high quality, coding variants in ID-associated genes from the merged VCF files. This list contained 9015 variants, of which 8476 were functional (8389 missense; 70 in-frame indels; and 17 variants resulting in loss of a stop codon) and 539 in total were LOF (221 nonsense; 189 frameshift; 77 essential splice donor; and 52 essential splice acceptor) (Figure 3-3). The average number of LOF variants per person was 0.54, while the average number of missense variants per person was 9.05.

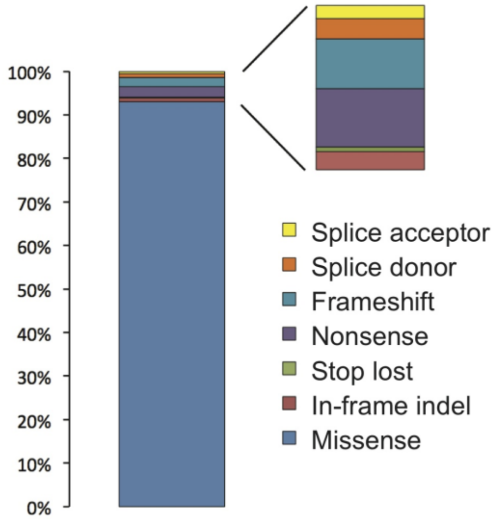

Figure 3 - 3: Classes of variant identified through the R filtering pipeline.
Total number of variants = 9015, total number of samples = 996.

Dr Lucy Raymond and Dr Detelina Grozeva imposed further stringent filters on this list of 9015 variants, to identify variants that are highly likely to be causative. These filters took into account factors such as the type of the variant, frequencies in the public and internal databases, presence in the human gene mutation database (http://www.hgmd.org/), consistency with the estimated mode of inheritance based on the Developmental Disorder Gene2Phenotype (DDG2P) gene list (https://decipher.sanger.ac.uk/), and the clinical phenotype of the affected individual. They validated a subset of the variants using either Sanger sequencing or exome sequencing of non-amplified DNA. Using this case-only diagnostic analytical approach, they found that 109 individuals (10.9%) had likely causative LOF variants, and 34 individuals (3.4%) had likely causative missense variants, giving a total estimated diagnostic yield of ~14%.

3.3.5 *SETD5* is a novel intellectual disability-associated gene

To identify novel ID-associated genes, Dr Lucy Raymond and Dr Detelina Grozeva focused on genes that had the highest number of LOF variants in the list that I generated. They found that seven individuals had a rare, high-quality, LOF variant in *SETD5* (0.7% of the cohort). They confirmed all the variants using Sanger sequencing, and confirmed that five are *de novo* by Sanger sequencing of parental DNA (paternal DNA was unavailable for two probands). I calculated that the probability of this occurring by chance in a cohort of this size is very low ($p = 5.25 \times 10^{-9}$). The mutations in *SETD5* were all different, and only one LOF mutation (which was more 3’ than any identified in these ID patients) was listed in the NIH Heart, Lung, and Blood Institute’s Exome Variant Server (NHLBI EVS). No other candidate gene was confirmed as being a novel ID-associated gene using this approach, because they either had a high number of LOF mutations listed in the NHLBI EVS database, or the variants were all the same, increasing the chances they are in fact a sequencing error (Table 3-4).

Gene	Total number LOFs	Number Independent LOFs	Number NHLBI EVS LOF	Reason excluded from further analysis
DCHS2	22	9	13	High number LOFs in NHLBI EVS
SETD5	7	7	1	NA
MIB1	9	7	13	High number LOFs in NHLBI EVS
STAB2	6	6	12	High number LOFs in NHLBI EVS
PCDH10	7	1	1	Low number of independent LOFs
UTP14A	6	1	0	Low number of independent LOFs

Table 3 - 4: Candidate genes with the highest number of LOF variants.
Table includes candidate genes not listed as ID-associated in OMIM. Table is sorted according to number of independent LOFs. Data courtesy of Dr Detelina Grozeva.

An international team of collaborating clinicians documented and compared the phenotypes of the seven patients with *SETD5* mutations. In addition to ID, there were several common and recurring features including ritualised behavior or ASD, abnormal ears, eyebrows, eyes, and nose, and skeletal and gastrointestinal abnormalities. They noticed that the facial appearance of the cases was, in some aspects, strikingly similar (Figure 3-4). Due to the phenotypic similarity of the cases, and the small probability of this many mutations occurring by chance, we concluded that these LOF mutations in *SETD5* are causative in these seven patients, and that LOF of *SETD5* causes a potentially recognisable syndrome. Indeed, LOF of *SETD5* may be a relatively common cause of ID (191).

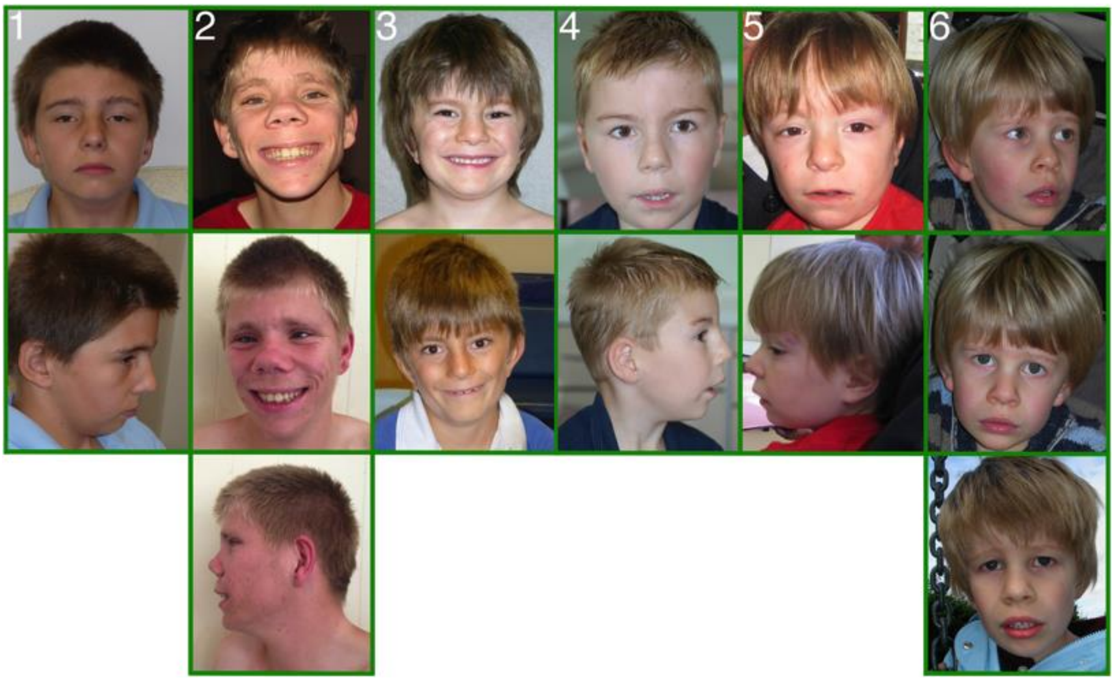

Figure 3 - 4: Facial appearance of individuals with *SETD5* mutations.
Photographs of the seventh patient were unavailable. This figure is courtesy of Dr Lucy Raymond, and it has been published (191).

3.3.6 Individuals with intellectual disability have an enrichment of loss of function variants in sequenced ID-associated genes, compared to controls

I used the CAST method to assess the extent to which LOF variants in sequenced ID-associated genes are enriched in the ID cohort compared to the CHD cohort. I selected CAST rather than one of the other methods such as the weighted sum method or

SKAT, because according to the DDG2P list the mechanism of the vast majority of known ID-associated genes is loss of or reduction of protein function, so I think that the vast majority of causative variants in this cohort will have the same direction of effect.

First I excluded samples that had an excessive number of LOF variants (>4, which is >3.5 standard deviations from the mean number of variants per sample). Of the 986 ID samples remaining, 341 (34.6%) had at least one rare (internal frequency <1%) LOF variant in a sequenced ID-associated gene, compared to 225/903 (24.9%) in CHD. This represents a highly significant enrichment (p = 2.8 x 10^{-6}) (Figure 3-5). This difference between the cohorts is most likely accounted for by the fraction of LOF variants that are causative of ID, suggesting that ~10% of ID cases in this cohort are caused by LOF variants in the sequenced genes. This is very consistent with the manual case-only diagnostic analysis, in which 109 (10.9%) cases were found to be caused by LOF variants.

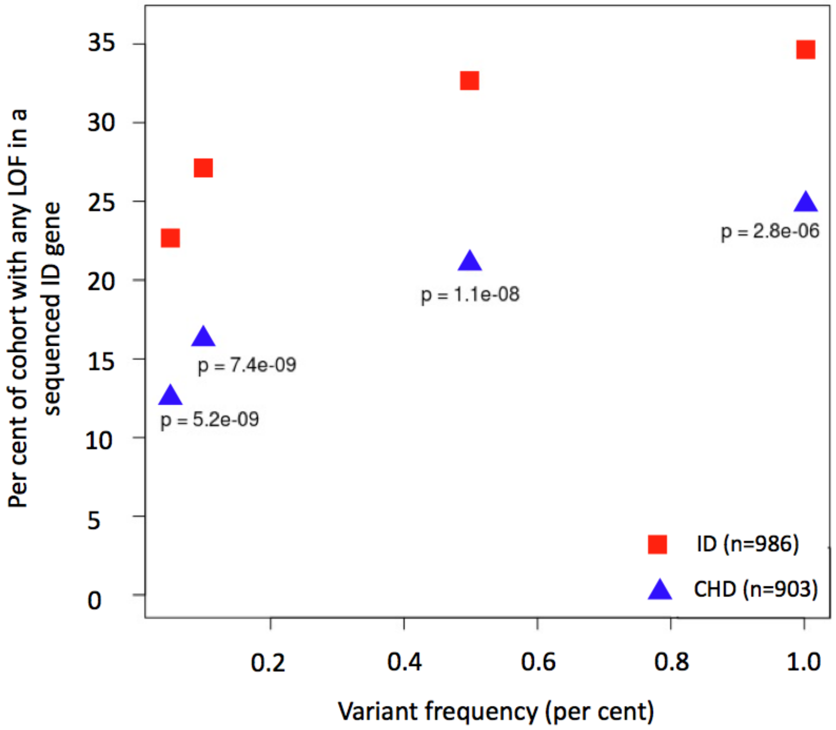

Figure 3 - 5: Patients with intellectual disability have an enrichment of loss of function variants in sequenced intellectual disability-associated genes compared to controls. LOF = loss of function; ID = intellectual disability; CHD = congenital heart disease. Numbers in key show number of samples. P values were calculated by one-tailed Fisher's exact test.

I next applied more stringent internal variant frequency filters of 0.5%, 0.1% and 0.05%, the latter of which leaves unique variants only. Of the 986 ID samples, 223 (22.6%) had at least one unique LOF variant in a sequenced ID-associated gene, compared to 113/903 (12.5%) in CHD. Therefore, after application of this more stringent filter, the difference of around 10 percentage points between the cohorts is maintained, and the enrichment of LOF variants in ID becomes more significant (p = 5.2 x 10^{-9}). This suggests that the vast majority of the LOF variants that cause ID in this cohort are unique within the cohort.

The LOF variants can be categorised according to chromosome, variant type, whether the sequenced ID-associated gene is known or a candidate, and whether it causes disease according to a biallelic or a non-biallelic mode of inheritance. I performed the CAST test to evaluate the degree of enrichment of each of these categories of unique LOF variants in the ID cohort (Table 3-5).

Gene category		**Variant type**	**Number LOFs ID**	**Number LOFs CHD**	**P-value**
Autosome or PAR	Known non-biallelic *76*	SNV	42/986 (4.26%)	8/903 (0.89%)	1.922 x 10^{-6}*
		Indels	14/986 (1.42%)	7/903 (0.78%)	0.132
	Known biallelic *52*	SNV	25/986 (2.54%)	16/903 (1.77%)	0.164
		Indels	15/986 (1.52%)	6/903 (0.66%)	0.058
	Candidate *212*	SNV	67/986 (6.8%)	32/903 (3.54%)	9.795 x 10^{-4}*
		Indels	33/986 (3.35%)	30/903 (3.32%)	0.54
X chromosome (males only)	Known *76*	SNV	13/925 (1.41%)	0/467 (0%)	0.0048*
		Indels	11/925 (1.19%)	0/467 (0%)	0.011
	Candidate *149*	SNV	14/925 (1.51%)	2/467 (0.43%)	0.056
		Indels	7/925 (0.76%)	1/467 (0.21%)	0.191

Table 3 - 5: Enrichment of unique LOF variants in the ID cohort, split by category.
The numerator in the ‘Number LOFs ID’ and ‘Number LOFs CHD’ columns show the number of samples in each cohort that have one of more unique LOF variant of the category indicated. The number of genes in each category is given in italics. PAR = pseudo-autosomal region; SNV = single nucleotide variant; LOF = loss of function variant, ID = intellectual disability cohort; CHD = congenital heart disease control cohort. P values calculated using Fisher’s exact test. *Below Bonferroni-corrected threshold of 0.005.

LOF SNVs in autosomal, known ID-associated genes with non-biallelic mode of inheritance are significantly enriched in the ID cohort ($p = 1.922 \times 10^{-6}$). In contrast, I identified no significant enrichment in known ID-associated genes with biallelic mode of inheritance ($p = 0.164$). Given that the parents of the probands in this cohort are unaffected, this suggests that dominant, *de novo* mutations are an important cause of disease in our cohort. This is consistent with studies showing that *de novo* LOF mutations are a particularly important cause of ID (57, 84).

Furthermore, LOF SNVs in autosomal, candidate ID-associated genes are significantly enriched in the ID cohort ($p = 9.795 \times 10^{-4}$). This very strongly suggests that some of these candidate genes are real ID-associated genes, even though they have not yet been definitively proved as such. Unfortunately, I could not use the CAST test to identify the individual candidate genes that were driving this signal, because relatively small cohort sizes and effect sizes render the CAST test underpowered for this purpose. Additionally, LOF SNVs in X-linked, known ID-associated genes in males are significantly enriched in the ID cohort ($p = 0.0048$). Interestingly, I identified no significant enrichment in X-linked candidate genes ($p = 0.056$). This suggests that, compared to the autosomes, a higher proportion of ID-associated genes on the X chromosome have been identified. This is unsurprising, as the X chromosome has been disproportionately well studied in ID (13). I did not detect any significant enrichment of LOF indels, which is likely due to reduced sensitivity of indel calling programs compared to SNVs.

There is no significant enrichment of synonymous variants in sequenced ID-associated genes in the ID cohort compared to CHD ($p = 0.475$). Subcategorising the synonymous variants reveals no significant enrichment in any category (data not shown). This is important because if the enrichment of missense variants was a spurious result due to a difference in the cohorts such as population stratification, one would expect to see an equivalent enrichment in synonymous variants. This finding therefore increases the chance that the observed enrichment is real and biologically relevant.

3.3.7 In known ID-associated genes on the X chromosome, unique missense variants tend to be more damaging in ID patients than controls.

To test the hypothesis that unique, missense variants in sequenced ID-associated genes are more likely to be damaging in the ID cohort than the CHD cohort, I compared the distribution of PolyPhen2, SIFT, and Condel scores using one-tailed, unpaired Mann-Whitney tests (65, 66, 196). I excluded samples with excessive numbers of missense variants (>25, which is >3.5 standard deviations from the mean number of variants per sample), and individuals in the ID cohort for whom a clearly causal LOF variant had been identified, from this analysis. For all scores, the only category of variant where there was a significant difference between the cohorts was missense variants in X-linked, known ID-associated genes. In this category, variants in ID cases were predicted to be significantly more damaging than those in controls ($p < 0.0001$) (Figure 3-6), suggesting that a proportion of this category of missense variant do indeed cause ID. In contrast, in known ID-associated genes on the autosomes, there is no difference in scores of predicted damage of unique missense variants between ID patients and controls (Figure 3-7).

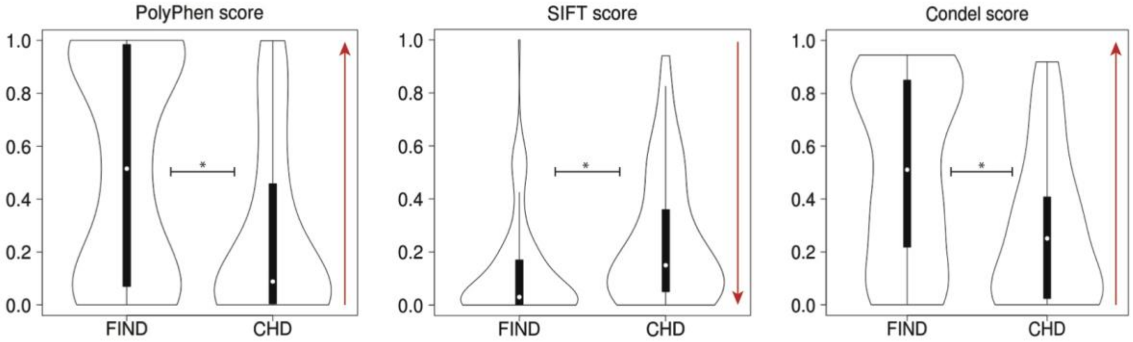

Figure 3 - 6: In known ID-associated genes on the X chromosome, unique missense variants are predicted to be more damaging in ID patients than controls.
The number of samples (ID = 825; CHD = 466) does not include those with excessive numbers of missense variants (>25), or ID samples with causative LOF variants identified. These plots consist of 154 missense variants in known ID-associated genes for the ID cohort, and 62 for the CHD cohort. * = p < 0.0001, calculated by Mann-Whitney tests. The red arrow on each plot indicates the direction of increase in predicted damage.

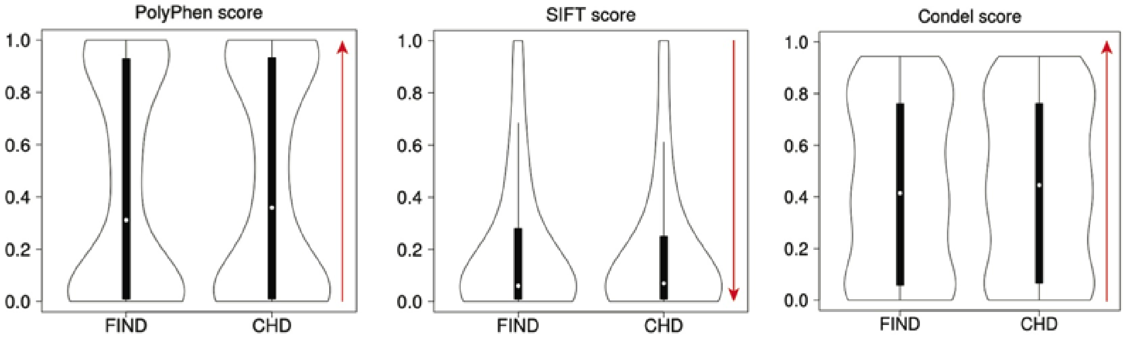

Figure 3 - 7: In known ID-associated genes on the autosomes, unique missense variants are not predicted to be more damaging in ID patients than controls.
The number of samples (ID = 877; CHD = 900) does not include those with excessive numbers of missense variants (>25), or ID samples with causative LOF variants identified. These plots consist of 1184 missense variants in known ID-associated genes for the ID cohort, and 1039 for the CHD cohort. There is no significant difference in scores of predicted damage between ID cases and controls (Mann-Whitney test, p > 0.66). The red arrow on each plot indicates the direction of increase in predicted damage.

3.3.8 Evidence for an enrichment of unique, predicted damaging, missense variants in sequenced ID-associated genes in the ID cohort

One reason that detecting an enrichment of missense variants in case-control analyses is harder than for LOF variants is that a smaller proportion of missense than LOF variants cause disease. Therefore, any enrichment of damaging missense variants in the ID cohort could be masked by the ‘noise’ of benign missense variants. I therefore applied the CAST test to unique missense variants that are predicted to be damaging by at least one of the three scores of predicted damage, in order to assess the possible contribution of causal missense variants in our cohort. I first excluded samples with excessive numbers of missense variants (>25). In the ID cohort, I also excluded the 109 samples for which a clearly causal LOF variant had been identified.

Of the ID samples, 691/877 (78.8%) had at least one unique, predicted damaging, missense variant in a sequenced ID-associated gene, compared to 688/900 (76.4%). This does not represent a significant enrichment (p = 0.129). However, two of the subcategories do have a significant enrichment (Table 3-6). Of the ID samples, 438/877 (49.9%) had at least one unique, predicted damaging, missense variant in a candidate autosomal gene compared to 393/900 (43.7%) CHD samples (p = 0.005), suggesting that around 6% of cases in our cohort might be caused by this category of variant. This suggests that variants in a subset of these candidate genes can indeed cause ID, which is consistent with the results of the CAST test on LOF variants. It is interesting that there is a more significant enrichment for candidate than known ID-associated genes. This could be a consequence of there being more candidate than known genes, or it could be that a higher proportion of candidate than known ID-associated genes operate by a non-LOF mechanism.

Gene category		**Number missense ID**	**Number missense CHD**	**P-value**
Autosome or PAR	Known non-biallelic *76*	258/877 (29.4%)	232/900 (25.8%)	0.048
	Known biallelic *52*	233/877 (26.6%)	213/900 (23.7%)	0.088
	Candidate *212*	438/877 (49.9%)	393/900 (43.7%)	0.005*
X chromosome (males only)	Known *76*	86/825 (10.4%)	17/466 (3.6%)	4.65 x 10^{-6}*
	Candidate *149*	169/825 (20.5%)	78/466 (16.7%)	0.057

Table 3 - 6: Enrichment of unique, predicted damaging, missense variants in the ID cohort, split by category.
The numerator in the ‘Number missense ID’ and ‘Number missense CHD’ columns show the number of samples in each cohort that have one or more unique, predicted damaging missense variant of the category indicated. The number of genes in each category is given in italics. The number of total samples does not include those with excessive numbers of missense variants (>25), or ID samples with causative variants identified. PAR = pseudo-autosomal region; ID = intellectual disability cohort; CHD = congenital heart disease control cohort. P values calculated using Fisher’s exact test. *Below Bonferroni-corrected threshold of 0.01.

Up to 7% of cases in our cohort might be caused by unique, predicted damaging, missense variants in known ID-associated genes on the X chromosome, because 86/825 (10.4%) males in the ID cohort have at least one, compared to 17/466 (3.6%) in the CHD cohort ($p = 4.65 \times 10^{-6}$).

3.4 Discussion

3.4.1 Summary

A targeted resequencing study was carried out as part of the UK10K project; 565 ID-associated genes were sequenced in 996 ID patients. I generated a list of rare, high quality, coding variants in the ID-associated genes in this cohort. From these data, causative variants were identified for ~14% of the cohort, and the novel ID-associated histone methyltransferase gene *SETD5* was identified. I next confirmed that there is no substantial difference in population structure between the ID cases and controls with CHD, and I used CAST to identify a highly significant enrichment of unique LOF variants in ID-associated genes in cases compared to controls. The size of the burden was consistent with the findings of the case-only diagnostic analysis. I subcategorised the LOF variants according to features of the variant itself, and features of the gene that it affects. From this, I found that the enrichment is greater in known than candidate genes, it is greater in genes with a non-biallelic rather than a biallelic mode of inheritance, and it is greater in SNVs than indels. I extended the analysis to missense variants. There was lower power to detect enrichment in missense variants, because a lower proportion of them are casual. Nevertheless, I found a moderately significant enrichment of missense variants in candidate autosomal genes, and a highly significant enrichment in known ID-associated genes on the X chromosome. This is consistent with the observation that missense variants in known ID-associated genes on the X chromosome are, on average, predicted to be significantly more damaging in ID cases than controls with CHD.

3.4.2 Loss-of-function of the histone methyltransferase gene *SETD5* is probably responsible for the cardinal features of 3p25 microdeletion syndrome

In this study, we showed for the first time that *de novo* LOF mutations in the histone methyltransferase gene *SETD5* cause ID, along with additional phenotypes such as ritualised behaviour, and dysmorphic facial features (191). In our cohort, this was a relatively frequent cause of disease, accounting for 0.7% of cases, which is similar to the frequency of *ARID1B* mutations, which are considered to be one of the more common causes of sporadic ID (173).

There are three reasons why *SETD5* was selected as a candidate ID-associated gene to be sequenced in this study. First, a *de novo* LOF mutation in *SETD5* was reported in a single ID patient in a previous study (84). While intriguing, this was not sufficient for Rauch *et al.* to conclude that *SETD5* is definitely an ID-associated gene, and the authors did not extensively report the phenotype of this patient. Second, *de novo SETD5* mutations have been associated with ASD in several studies, and it is widely known that there is much overlap in the presentation and genetic aetiology of ID and ASD (197-199). Third, *SETD5* is one of only two protein-coding genes in the minimal critical region for 3p25 microdeletion syndrome (200).

The 3p25 microdeletion syndrome was first described in 1978 (201). Since then there were several other case reports of *de novo* deletions at this locus, resulting in phenotypes including ID, seizures, microcephaly, CHD, malformed ears and nose, and other dysmorphic craniofacial features (202-204). The sizes and breakpoints of the deletions in these cases varied, and so the minimum critical region was refined over time. Most recently, a case report refined it to only 124 kb, containing only three genes: *THUMPD3*, *SETD5*, and *LOC440944* (an RNA gene) (200).

The phenotypes of the patients with *SETD5* mutations described in this study are very similar to those of the patients with 3p25 microdeletion syndrome (191). Phenotypes that overlap in both groups include ID, abnormal eyebrows, a depressed nasal bridge, large or low-set ears, a long smooth philtrum, OCD or ritualised behaviour, skeletal abnormalities, and CHD. With the exception of ID, these phenotypes are variable, appearing in multiple, but not all, cases. The overlap between the two groups is not complete; for example, none of the patients in our study had seizures or microcephaly, which are features of some cases of 3p25 microdeletion syndrome. Therefore, while the possibility that haploinsufficiency of 3p25 genes other than *SETD5* might contribute to the clinical phenotype in some patients cannot be excluded, it appears highly likely that haploinsufficiency of *SETD5* is responsible for the cardinal features of 3p25 microdeletion syndrome.

One study of CNVs in patients with ASD came to a different conclusion. Pinto *et al.* identified a 24 kb deletion encompassing most of *SETD5* and no other genes in a single patient with ASD and borderline ID, but no other medical issues or dysmorphic features (199). They therefore suggest that while LOF of *SETD5* may be at least partially responsible for the intellectual and behavioural deficits of 3p25 microdeletion syndrome patients, it is probably not involved in the other features of the syndrome.

Pinto *et al.* was published before the *SETD5* study, so the authors were unaware of the seven patients described here (191). It is more likely that some form of genetic compensation explains the mild phenotype in their single patient, than that the overlapping phenotypes in the seven patients with *SETD5* mutations in this study are coincidental.

Interestingly, ID caused by *SETD5* mutations is another example of a clearly syndromic form of ID that is not recognised as such until a group of patients with shared aetiology are retrospectively examined together. This emphasises the importance of assembling groups of patients with shared aetiology. *SETD5* can also be added to the list of ID-associated genes that were discovered after being identified as candidates because they are in a CNV. Historically, this has been an important way to identify ID-associated genes, particularly in autosomes. Other ID-associated genes that were identified this way include *MBD5* and *KANSL1* (156, 157).

Before describing variants in a gene as causative of any rare disease, it is important to apply a high and consistent standard to the evidence assembled to support the assertion. For example, a rare variant that segregates with Mendelian disease in a single family is not necessarily causative (12). As sample sizes and the amount of sequencing data increases, the probability of finding recurrent similar variants in a given gene just by chance also increases. Therefore, it is also important to apply statistical tests to demonstrate that the variants in question are significantly enriched in patients. Furthermore, if LOF variants in a given gene are relatively common in the general population it is unlikely that LOF of that gene causes a rare disease. Several ID-associated genes have recently been called into question on this basis (64). Therefore in this study, my colleagues and I took care to apply a high standard of evidence to the data, before concluding that *SETD5* is a novel ID-associated gene. For example, we showed that LOF of *SETD5* in the general population is very rare, and we showed that the mutations identified were highly unlikely to have occurred by chance (191).

SETD5 is predicted on the basis of sequence homology to encode a histone methyltransferase (205). As well as *SETD5* and *EHMT1* (pathogenic variants in which can cause Kleefstra syndrome as discussed) known ID-associated histone methyltransferases include *EZH2* and *MLL2* (also known as *KMT2D*). EZH2 is part of a complex that methylates a specific lysine residue on histone H3 (206). It has many important roles in development, including X chromosome inactivation, and stem cell regulation (207, 208). *De novo* mutations in *EZH2* can cause Weaver syndrome, features of which include ID, overgrowth, and characteristic craniofacial dysmorphic features (209). *MLL2*, mutations in which can cause Kabuki syndrome, which also involves ID, and also catalyses methylation of histone lysine residues (210). Therefore, although little more is known about the function of *SETD5*, histone methyltransferases are clearly emerging as a very important class of ID-associated genes. *SETD5* fits well into the known pattern for ID-associated histone methyltransferases, because all known causative mutations are *de novo*, and the resulting phenotype is syndromic. These two features are consistent with all the other known examples of ID-associated histone methyltransferases discussed.

3.4.3 Insights from case-control enrichment analyses

The case-control enrichment analysis demonstrates that in this cohort, 10% of ID cases are caused by LOF variants in the sequenced genes. This is consistent with the results of the case-only diagnostic analysis, in which a causative LOF variant was identified for 10.9% of the cohort. Using case-control enrichment analysis I estimate that up to 13% of cases in this cohort are caused by unique, predicted damaging, missense variants (6% in candidate autosomal genes, plus 7% in known ID-associated genes on the X chromosome). This is much higher than the rate of causative missense variants found by manual case-only diagnostic analysis, which is only 3.6%. This suggests that the true proportion of the cohort where disease is caused by missense variants is higher than 3.6%. However, assigning pathogenicity to missense variants with a diagnostic level of confidence is more difficult than for LOF variants, and must be done conservatively.

Two previous exome sequencing studies of ID cohorts have estimated diagnostic yields of 16% and 31% respectively (57, 84). Another exome sequencing study of children with developmental disorders, many of whom had ID, had a diagnostic yield of 25% (11). Differences in ascertainment and methodology make direct comparisons of diagnostic yield between studies problematic. There are four reasons why our total estimated diagnostic yield of 14% is lower than that of the previous studies. First, we resequenced the exons of 565 known and candidate ID-associated genes only in a targeted approach, rather than sequencing all genes. Second, we sequenced probands only, not trios. Third, 94% of this UK10K ID cohort is male, whereas most other cohorts are approximately 50% male, and it is possible that, on average, males with ID have a higher contribution from oligogenic causes. Finally, a proportion of cases in our cohort had been through extensive previous investigation, so the cohort is enriched for harder to solve cases.

Assigning causality to a novel candidate gene requires a high degree of evidence (12). In this study, the sizes of the cohorts were insufficient to have power to detect a significant enrichment of variants in individual novel candidate genes using case-control enrichment analyses. Nevertheless, the finding that there is a significant enrichment of both LOF and missense variants in candidate ID-associated genes shows that some of these variants must be causative. The enrichment of both LOF and missense variants in known genes with a non-biallelic mode of inheritance is greater than that in known genes with a biallelic mode of inheritance, which tells us that *de novo* mutations are probably an important cause of ID in our cohort, even though we did not sequence trios. These insights into the genetic architecture of the cohort highlight the utility of case-control enrichment analyses as a supplementary tool to manual case-only diagnostic analysis.

Fundamental differences between the X chromosome and autosomes may explain why the burden of missense variants is so much larger for known, ID-associated genes on the X chromosome, than for any other category of missense variants in this study. For example, a higher proportion of X chromosome genes are involved in brain development and function than autosomal genes (211-213). Given that this UK10K ID cohort is 94% male, one might therefore expect a disproportionate number of cases to be caused by pathogenic variants in the X chromosome because of this functional bias. Additionally, ID-associated genes have also been particularly well studied on the X chromosome, so a higher proportion of X-linked than autosomal ID-associated genes may have been identified (13).

Furthermore, differences between the X chromosome and autosomes may influence scores of predicted damage. Greater selection pressure acting upon the X chromosome results in less diversity on the X chromosome than autosomes (214). This also means that, in general, X chromosome genes are more conserved between species than autosomal genes (215). SIFT, for example, assesses how likely a variant is to be damaging, according to how conserved the affected locus is, with more conserved positions likely to be less tolerant to variation (66). As X chromosome genes are generally more conserved than autosomal genes, scores of predicted damage might be, on average, higher on the X chromosome than autosomes.

Unlike classic case-only diagnostic analysis, case-control enrichment analysis takes into account variants with incomplete penetrance, and variants that contribute to a phenotype in an oligogenic manner. However, when a burden of variants is identified, it is not currently feasible to distinguish how much of the burden is caused by causative variants with complete penetrance, and how much is caused by variants with incomplete penetrance, oligogenic variants, and secondary modifiers of phenotype. Purcell *et al.* described the burden that they identified in schizophrenia candidate genes as “polygenic”, but they use the term on the population level, and do not suggest that individuals necessarily have multiple causative alleles (14). Development of statistical methods that can distinguish between these scenarios would be a very welcome future development.

3.4.4 Limitations of this study

The major limitation of the study design is that we employed an inherently biased, targeted gene approach, in which only 565 known and candidate ID-associated genes were sequenced. This decision was taken for financial reasons, and it meant that any causative variants in other genes could not be identified, so the diagnostic yield is almost certainly lower than what it would have been had we done exome sequencing instead, for example. Similarly, only probands were included in this study, meaning that without performing additional sequencing, *de novo* mutations could not be distinguished from inherited variants, making it more difficult to interpret the results. The list of 565 known and candidate ID-associated genes was originally compiled in 2012, so now the list is quite out of date as many additional ID-associated genes have been identified since then (62).

Regarding the case-control enrichment analysis, the result that there is no enrichment of indels in ID cases compared to controls suggests that indels are called with low sensitivity by the UK10K pipeline. Another limitation to bear in mind is that categorisation of the sequenced ID-associated genes into known and candidate genes is to some extent a false dichotomy. This is actually a complex task, and the level of stringency required to distinguish between the two categories is not something on which the ID research community has reached a clear consensus. For example, some think that variants in a certain minimum number of unrelated cases must be identified before a gene can be classified as "known", as opposed to "candidate", whereas others do not think this is always necessary ((62) and personal communication from Dr Matthew Hurles). Similarly, there are genes such as *NF1* in which variants are associated with ID in a proportion of cases, but not in a high enough proportion of cases to be definitively classified as ID-associated genes (personal communication from Dr. Lucy Raymond). I decided to use a recently generated, manually curated, stringent list of known genes for this study (62). It is likely that some researchers would argue that some of the genes I have categories as "known" are actually "candidate", and vice versa.

Finally, it is disappointing that this study was underpowered to detect an enrichment of variants in individual genes. However, it is not at all surprising, as it has previously been shown that much larger samples sizes than ~1000 cases and ~1000 controls would be required to achieve this (190).

3.4.5 Further work

An ongoing project that will extend the work described in this chapter is a large exome sequencing project of 1151 individuals with ID or their relatives. Importantly, 541 (47%) of these individuals were also included in the targeted resequencing study described here. Therefore, the exome sequencing study will enable us to validate findings of the targeted resequencing study, and hopefully identify more causative variants and more novel ID-associated genes. This exome sequencing study includes several different family structures such as 49 trios and 121 affected sibling pairs, which will facilitate easier interpretation of variants than single probands too, because for example *de novo* or shared variants can be identified. At the time of writing, the sequencing, mapping, variant calling, and filtering for this study has been completed, and the data are being further analysed and interpreted.

Another exciting ongoing project is the development of a *SETD5* mouse model (https://www.komp.org/geneinfo.php?MGI_Number=1920145). Dr Jacqui White of the mouse genetics programme at WTSI has led this work. Homozygous null mice are unsurprisingly lethal, but early phenotyping on a small number of heterozygous mice so far suggests that they may have interesting features, such as dysmorphic craniofacial features, including a depressed nasal bone (personal communication from Dr Jacqui

White). This appears to confirm that *SETD5* has a role in development of the mid-face and the skull, as suggested by the seven patients in our study.

Plans are underway to assess the cognitive abilities of these mice. If the mice are indeed cognitively impaired, they could be valuable experimental tools with which to identify any downstream genes whose expression is altered as a result of LOF of *SETD5*. This might be achieved, for example, by performing RNA-seq on brain tissue from heterozygous *SETD5* knockout mice, along with their wildtype siblings as controls. This might really start to demonstrate how *SETD5* mutations cause ID, and could even ultimately lead to identification of therapeutic targets.

The most important outcomes of the work described in this chapter are as follows. We have identified a genetic diagnosis for ~14% of the ID patients in this UK10K cohort. We have identified *SETD5* as a novel ID-associated gene, supporting the importance of histone methyltransferases in the aetiology of ID. Additionally, we have demonstrated that LOF of *SETD5* is probably responsible for the cardinal features of 3p25 microdeletion syndrome. Finally, certain categories of variants are enriched in ID-associated genes in ID cases compared to controls, yielding insights into the genetic architecture of ID, and demonstrating the utility of case-control enrichment analysis as a supplementary analytical approach in large genomic studies of rare disease.

4 Modelling dystroglycanopathy in zebrafish embryos by knockdown of *B3GALNT2* and *GMPPB*

4.1 Introduction

4.1.1 The phenotypic spectrum of dystroglycanopathy

Dystroglycanopathy is a subtype of congenital muscular dystrophy (CMD) that is characterised by hypoglycosylation of the α-dystroglycan (α-DG) protein. Dystroglycanopathy is currently clinically classified according to phenotypic severity. Walker-Warburg Syndrome (WWS) is the most severe type of dystroglycanopathy. Affected individuals have such severe muscular dystrophy that they have essentially no muscle tone from birth. There is usually profound intellectual disability (ID), structural brain abnormalities (such as cobblestone lissencephaly, hydrocephalus, and cerebellar malformations), and eye involvement (such as retinal malformations, micropthalmia, and blindness) (216). WWS is usually not compatible with life beyond one year of age.

The slightly less severe subtype is muscle-eye-brain disease (MEB). The phenotype is essentially similar, although patients might live for a few years, and develop some limited communication skills and motor control (217). Online Mendelian Inheritance in Man (OMIM) has recently collectively termed WWS and MEB as muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies) type A1 (MDDGA1; MIM 236670).

In the intermediate range of the spectrum of phenotypic severity, dystroglycanopathy patients have moderate to severe CMD manifesting in infancy or childhood. They may or may not have ID and central nervous system (CNS) involvement (218), and eye abnormalities are rare. This form is called muscular dystrophy-dystroglycanopathy (congenital with mental retardation), type B1 (MDDGB1; MIM 613155).

The mildest form of dystroglycanopathy is limb girdle muscular dystrophy (LGMD). LGMD is characterised by relatively mild muscular dystrophy with later onset. There may or may not be ID or mild brain abnormalities (219). This form is known as muscular dystrophy-dystroglycanopathy (limb-girdle), type C1 (MDDGC1; MIM 609308) to distinguish it from LGMD patients who do not have hypoglycosylation of α-DG.

4.1.2 Dystroglycan structure, function and glycosylation

Both α-DG and β-dystroglycan (β-DG) are members of the dystrophin-glycoprotein complex (DGC), and they are encoded by a single gene, *DAG1* (MIM 128239). The dystroglycan polypeptide is post-translationally cleaved into the α and β subunits (220). β-DG spans the membrane of the sarcolemma. Intracellularly, it associates with dystrophin to transduce force from the myocyte protein machinery such as filamentous actin. Extracellularly, β-DG remains non-covalently bound to α-DG. α-DG in turn binds to extracellular matrix (ECM) proteins including laminin 2, agrin, perlecan, neurexin and pikachurin (216, 220-222).

α-DG is extensively glycosylated (Figure 4-1). There are five types of protein glycosylation: N-glycosylation, O-mannosylation, C-mannosylation, phospho-serine glycosylation and Glycosylphosphatidylinositol (GPI) anchor formation. α-DG undergoes O-mannosylation, and N-glycosylation, with O-mannosylation being most functionally important (223). First *O*-Mannose glycan residues are added to serine and threonine residues of the core α-DG protein, in the endoplasmic reticulum (ER) (224). The O-mannose next undergoes extension and branching to create complex glycan chains, the exact structures of most of which are not yet well characterised (223). α-DG migrates to the Golgi apparatus where N-acetylgalactosamine (GalNAc) initiated glycans are also added (225).

Binding of α-DG to its ligands is absolutely dependent upon appropriate glycosylation (226). Therefore, pathogenic variants in genes encoding enzymes involved in glycosylation can cause the hypoglycosylation of α-DG, which results in defective ligand binding, and the degeneration of muscle structure characteristic of dystroglycanopathy (227). The IIH6 antibody recognises O-mannosyl glycans, and can therefore be used experimentally to assess the level of glycosylation of α-DG in tissues.

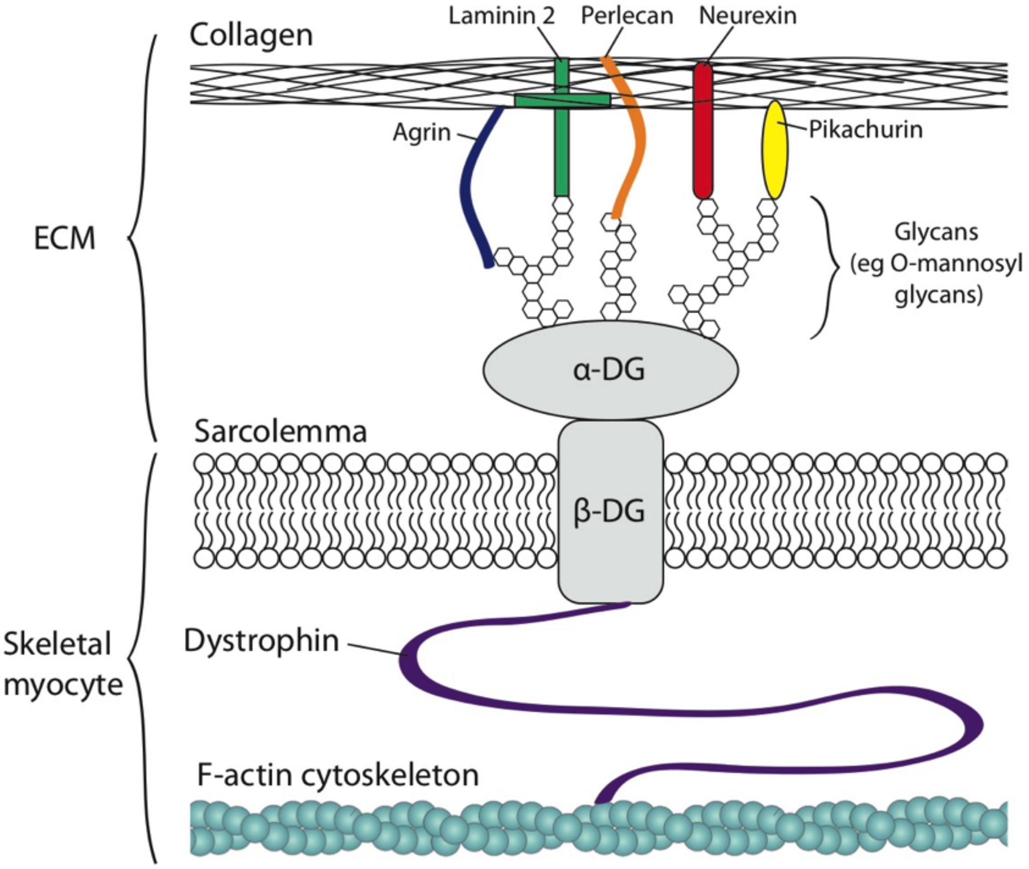

Figure 4 - 1: Model of α-DG interactions.
α-DG is non-covalently bound to the transmembrane protein β-DG, which connects to the sarcolemma cytoskeleton via dystrophin. α-DG binds to several components of the ECM via post-translationally added glycan chains, most importantly O-mannose glycans. Pathogenic variants in genes that encode proteins involved in this glycosylation cause impaired binding and loss of integrity of muscle tissue.

4.1.3 Known dystroglycanopathy-associated genes

Many proteins are involved in the glycosylation of dystroglycan, so it comes as no surprise that dystroglycanopathy is genetically as well as phenotypically heterogeneous. To date, not including the two genes described in this chapter, sixteen genes have been associated with dystroglycanopathy. First, there are rare reports of primary dystroglycanopathy, that is, dystroglycanopathy caused by pathogenic variants in the *DAG1* gene itself. The other fifteen genes are associated with secondary dystroglycanopathy. That is, pathogenic variants result in aberrant function of α-DG because of defects in the post-translational modifications of α-DG, not defects to the core structure of dystroglycan itself. These genes have autosomal recessive inheritance, and they can be classified into four further groups according to the function of the protein they encode. These groups are: proteins involved in synthesis of dolichol phosphate mannose (Dol-P-Man), proteins involved in O-mannosylation, glycosyltransferases, and proteins with currently unknown function.

Primary dystroglycanopathy (DAG1)

The first report of a possibly pathogenic variant in *DAG1* itself was made in 2010 (228). The patient had learning difficulties, white matter abnormalities, elevated creatine kinase (CK), dyspraxia and facial hypotonia but otherwise no muscle dysfunction. The authors describe these symptoms as a subset of the classical dystroglycanopathy phenotype. The patient had a *de novo*, 2 Megabase (Mb) deletion that overlapped *DAG1*. However, the mutation was heterozygous, whereas variants that cause secondary dystroglycanopathy are recessive. While interesting, because of this, and the atypical spectrum of phenotypes, this case is not regarded as clearly being a case of primary dystroglycanopathy.

A year later, homozygous missense variants in *DAG1* were identified in a patient with LGMD, cognitive impairment, and hypoglycosylation of α-DG (229). This was a more typical dystroglycanopathy phenotype and genotype. A mouse model with the patient's variant had similar abnormalities. Despite this, some scientists in the dystroglycanopathy research community were skeptical that this variant was the sole cause of the patient's phenotype, because it was a single patient and the variants were missense. This skepticism was somewhat allayed by two subsequent studies. In one, simulations suggest that docking between α-DG and its ligands was weakened by the variant in that patient (15). In the other, a homozygous missense variant in *DAG1* was found in two siblings who had severe MEB with macrocephaly and white matter disease (230). The variant, which is in a conserved section of the part of *DAG1* that encodes β-DG, is thought to disrupt the structure of the protein.

Genes encoding proteins involved in synthesis of Dol-P-Man (DPM1, DPM2, DPM3, and DOLK)

DPM synthase is an enzyme which catalyses the formation of Dol-P-Man, a mannosyl donor. In mammals, DPM synthase is a complex consisting of three subunits: DPM1, DPM2 and DPM3 (231). DPM1 is the primary catalytic subunit. While DPM2 may have some enzymatic activity, its primary function is to stabilise DPM3, which in turn allows DPM1 to be stably expressed at the ER membrane (232).

Defects in the genes encoding the components of DPM synthase can cause wide ranging phenotypic effects. For example, pathogenic variants in *DPM1* can cause congenital disorder of glycosylation type Ie (CDG-Ie). These patients have intellectual and motor disability caused by defects in N-linked glycosylation of proteins (233). In 2009, pathogenic variants in *DPM3* that cause dystroglycanopathy were found (234). These patients had reduced O-mannosylation of α-DG. Pathogenic variants in *DPM2* and *DPM1* have since also been found in dystroglycanopathy patients (235, 236). This demonstrates that the two phenotypically distinct conditions congenital disorder of glycosylation (CDG) and dystroglycanopathy can have the same molecular cause. This is because Dol-P-Man is required for both N-glycosylation and O-mannosylation. It is unclear why different variants in this complex cause the two different phenotypes.

Similarly, DOLK catalyses the formation of dolichol monophosphate, which is a precursor of Dol-P-Man. Variants in *DOLK* have been reported in patients with a phenotype that overlaps dystroglycanopathy and CDG, with defective N-glycosylation and reduced O-mannosylation (237).

Genes encoding proteins involved in O-mannosylation (POMGnT1, POMT1, and POMT2)

POMGnT1 catalyses the transfer of N-acetylglucosamine (GlcNAc) to O-mannose, extending *O*-mannosyl glycan chains (217). POMT1 and POMT2 each have O-mannosyltransferase activity, but this depends on them interacting physically and functionally (238, 239). However, the functions of POMT1 and POMT2 are not interchangeable (240).

In 2001, pathogenic variants in *POMGnT1* were first found to cause MEB in six dystroglycanopathy patients (217). Since then, the phenotypic spectrum of patients with *POMGnT1* variants has been expanded, as some very mildly affected patients have been identified, as well as severely affected patients (241). Different classes of pathogenic *POMGnT1* variants have been implicated, including a duplication in the promoter and an intragenic deletion (242, 243). Patients with pathogenic *POMGnT1* variants have hypoglycosylated α-DG that has reduced ability to bind to its ligands (226).

In 2002, pathogenic *POMT1* variants were found to cause WWS with hypoglycosylation of α-DG in five consanguineous families (216). They can also cause a milder LGMD phenotype, and in some cases cardiac defects (244, 245). In 2005, *POMT2* variants were also implicated in dystroglycanopathy (221). A variety of pathogenic *POMT2* variants can cause disease including intronic deletions and substitutions (246). Pathogenic variants in *POMT1* or *POMT2* that are found in dystroglycanopathy patients can ameliorate catalytic activity (247).

Genes encoding glycosyltransferases (LARGE, B3GNT1, GTDC2, and POMK)

LARGE is a ubiquitously expressed glycosyltransferase that interacts in the Golgi apparatus with α-DG domains, including the N-terminal domain and the mucin-like domain. LARGE is required for glycosylation of these domains (248). Specifically, LARGE catalyses the addition of repeating units of xylose and glucaronic acid to α-DG, which is required for α-DG to bind to its ligands including laminin (249). Pathogenic variants in *LARGE* can cause dystroglycanopathy (250). Many different *LARGE* variants have been implicated, including intragenic rearrangements (243). Some patients' variants perturb the interaction of LARGE with α-DG (251).

B3GNT1 is an enzyme that interacts with LARGE, and catalyses polymerization of GlcNAc residues (252, 253). B3GNT1 is necessary for α-DG glycosylation (222). In 2013, pathogenic variants in *B3GNT1* were found to cause dystroglycanopathy (16). This study also showed that in human cells, wildtype B3GNT1 increases α-DG glycosylation, but the variant form of B3GNT1 does not.

In 2012, a combination of homozygosity analysis and WES in consanguineous families with WWS was used to find pathogenic variants in *GTDC2* (254). GTDC2 catalyses the addition of GlcNAc epitopes to O-mannosylated α-DG in the ER (255). Pathogenic variants reduce the catalytic activity of GTDC2. Pathogenic variants in *POMK*, a glycosyltransferase also known as *SGK196*, can also cause dystroglycanopathy (256, 257). POMK, along with GTDC2, is involved in synthesising an O-mannosyl trisaccharide structure on α-DG, without which α-DG cannot bind to laminin (258).

Genes encoding proteins with unknown function (FKTN, FKRP, ISPD, and TMEM5)

FKTN was the first identified dystroglycanopathy-associated gene. There is a particular subtype of dystroglycanopathy known as Fukuyama-type CMD (FCMD) which is relatively common (~1/10,000 births) in Japan. The phenotype consists of hypotonia and muscle weakness first manifesting in infancy, motor delay, ID, seizures in around half of patients, and malformations of the eye and brain (259). In 1998, a 3 kilobase (kb) retrotransposal insertion of a tandem repeat in the 3' untranslated region (UTR) of a novel gene they called *fukutin* or *FKTN*, was identified as the cause of disease in nearly 90% of FCMD cases (260). It is now thought to be a founder mutation that arose around 2000 years ago (261). The mutation results in inappropriate splicing of mRNA (262). Since this discovery, many additional *FKTN* variants that can cause dystroglycanopathy have been identified, both inside and outside the Japanese population (263, 264). Pathogenic *FKTN* variants have also been shown to cause a wider range of severity of phenotypes than just classical FCMD, from LGMD to WWS (265-267). α-DG is hypoglycosylated in the muscle of FCMD patients, and therefore less able to bind to its ligands including laminin, neurexin or agrin (226, 268).

In 2001, pathogenic variants in the *FKTN* related gene *FKRP* were found to cause severe CMD known then as MDC1C (269). This was closely followed by the finding that *FKRP* is also associated with a milder LGMD phenotype with reduced glycosylation of α-DG (270, 271). Pathogenic variants in *FKRP* can also cause a more severe phenotype with cardiac and CNS involvement (272). Both FKTN and FKRP are required for α-DG glycosylation, and they are predicted to be glycosyltransferases. However, despite years of research, their precise functions remain elusive.

Nevertheless, insights can be gained by examining tissue expression patterns and cellular localisation. *FKTN* is expressed prenatally in the brain, particularly in glial cells and astrocytes. This expression is reduced in FCMD cases (273, 274). The subcellular localisation of wildtype FKTN is the Golgi apparatus (275). Some CMD patients' pathogenic variants cause FKTN to be retained in the ER, probably because it is misfolded (276). Interestingly, FKTN interacts with POMGnT1, possibly modulating its activity (277).

In contrast, the subcellular localisation of FKRP has proved controversial. One study found that wildtype FKRP localises to the rough ER, and is therefore likely to be involved in an earlier stage of α-DG glycosylation pathway than FKTN (275). Two other studies contradicted this however. One found FKRP colocalised with the DGC in the sarcolemma (278), whereas the other concluded that wildtype FKRP is located in the Golgi apparatus, and that some pathogenic variants cause FKRP to be retained and degraded in the ER (279). A further study could not replicate this finding that wildtype and variant FKRP are differently localised (280).

Pathogenic variants in the uncharacterised gene *ISPD* can disrupt O-mannosylation of α-DG, causing dystroglycanopathy (281, 282). As for many dystroglycanopathy-associated genes, the phenotypic spectrum is variable (283). In cultured patients' fibroblasts, wildtype ISPD can restore glycosylation of α-DG (281). Pathogenic variants in *ISPD* and another uncharacterised gene *TMEM5* have been found in fetuses with cobblestone lissencephaly (284). This brain malformation often occurs in patients with dystroglycanopathy. *TMEM5*, which has a predicted glycosyltransferase domain, was confirmed as a dystroglycanopathy-associated gene in 2013 (256).

4.1.4 Frequency of variants, and genotype-phenotype correlations

Several groups have studied the frequency of pathogenic variants in the different dystroglycanopathy-associated genes. Pathogenic variants in *POMT1*, *POMT2*, *POMGnT1*, *FKRP*, *FKTN* and *LARGE* collectively account for up to 50% of cases of dystroglycanopathy, depending on ethnicity, and how the cohort has been screened (219, 285-287). These studies largely agree that *POMT1* and *POMT2* are the most frequently implicated genes, followed by *POMGnT1* and *FKRP*. The most recent of these studies was published in 2009. It is likely that pathogenic variants in all of the currently known dystroglycanopathy-associated genes, including those identified since 2009, will account for more than 50% of cases.

One might assume that pathogenic variants in different dystroglycanopathy-associated genes might cause slightly different phenotypes. While it often initially appears that pathogenic variants in one gene cause a particular subtype of dystroglycanopathy (e.g. *FKTN* variants cause FCMD, variants in *POMGnT1* cause MEB, and variants in *POMT2* cause WWS (217, 221, 260)) further study of each of these cases invariably reveals that the phenotypic spectrum is in fact broad and overlapping (267, 288, 289). In general, there appears to be remarkably little correlation between genotype and phenotype, although some groups have noted that pathogenic *POMT1* and *POMT2* variants tend to cause CNS involvement more frequently than variants in other genes, pathogenic *POMGnT1* variants often cause cerebellar cysts, and loss of function variants may cause more severe phenotypes than missense variants (219, 287, 290, 291). The general lack or correlation between genotype and phenotype make it virtually impossible to predict the causative gene from the phenotype alone.

Even more surprisingly, there appears to be little correlation between the severity of phenotype and the extent of hypoglycosylation of α-DG. For example, one study found that while there was some evidence of such a correlation for a few genes, in other cases, patients could have very mild phenotypes despite a profound hypoglycosylation of α-DG (292). One hypothesis that might explain this observation is that some dystroglycanopathy-associated genes may be involved in the glycosylation of other target proteins in addition to α-DG, and this could contribute to the severity of the phenotype. In another study, the severity of the phenotype in patients with pathogenic *POMGnT1* variants does not correlate with the levels of POMGnT1 protein (293).

4.1.5 Zebrafish models of genetic disease

In vivo models have for decades been vital tools in the study of human diseases. Mice represent the classic laboratory model for human diseases. However other organisms, including zebrafish, are increasing in popularity as disease models for many reasons. Zebrafish are vertebrates with a high homology to humans; 71.4% of human protein-coding genes have at least one direct orthologue in the zebrafish (294). This figure increases to 82% when only looking at genes associated with human disease. An important advantages of zebrafish over mice is that because the embryos develop *ex utero*, it is possible in some cases to use zebrafish to study the effect of knocking out a gene that is embryonic lethal in a placental mammal such as a mouse (one example of this is *dag1*) (295, 296). From a practical point of view, zebrafish are relatively cheap and easy to maintain, they have high fecundity, the embryos develop quickly, and they are amenable to various forms of genetic manipulation.

Zebrafish embryos are particularly good models with which to study muscle diseases, including dystroglycanopathy. In part, this is because a high proportion of each zebrafish embryo is muscle tissue, and the embryos are optically transparent, allowing easy visualization of developing muscle tissue using simple microscopy techniques. Conveniently, the muscle tissue develops quickly; the embryo can swim by 24 hours post fertilisation (hpf), and the muscle is fully differentiated by 48 hpf (295). Furthermore, despite some differences between human and zebrafish muscle tissue at the structural level, they are highly orthologous at the molecular level.

Disadvantages of zebrafish compared to mice include the fact that obviously they are more evolutionarily diverged from humans, and there are some structures, such as the lungs, hair, and teeth, which they do not have, while some organs are fundamentally different, such as their hearts which only have two chambers. However, it is possible to recapitulate genetic cardiac abnormalities in zebrafish hearts despite this profound difference between species (297). Perhaps surprisingly, there are even some early hints that complicated social behaviours in humans have some parallels in zebrafish (298).

There are three main methods of making a zebrafish model of disease: random mutagenesis, site-specific nucleases, and transient inhibition of gene expression. Random mutagenesis involves exposing zebrafish to a mutagen such as *N*-ethyl-*N*-nitrosourea (ENU), which generates random mutations in the genome that are propagated to the next generation. The aim of the zebrafish mutation project is to use this method to generate a zebrafish knockout for every protein-coding gene in the zebrafish genome (299). Random mutagenesis is an efficient and high-throughput method of generating mutants, but identifying all the mutations, and linking them to phenotypes, is challenging. The random nature of the mutations generated also limits the utility of this approach.

Site-specific nucleases are enzymes that consist of two parts: a DNA recognition motif to target the enzyme to a particular genomic locus, and a nuclease to induce double-strand breaks at that locus. Examples include zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPRs) (300). These enable very specific mutations to be introduced into the genome. ZFNs were the first to be developed, but they were not widely used because they were difficult, expensive and time-consuming to make. TALENs and CRISPRs are more promising tools, which can be made in individual laboratories relatively cheaply and easily, and in a high-throughout manner. Problems with site-specific nucleases include possible off-target effects, and the fact that the time from designing the system to having a mutant fish ready to phenotype takes at least several months (301).

The most commonly used tool with which to transiently inhibit expression of a zebrafish gene is morpholino oligonucleotides (MOs). These short, synthetic oligonucleotides are designed to specifically target the mRNA of a gene of interest, and prevent it from being properly spliced or translated. The main advantage of MOs over random mutagenesis or site-specific nucleases is that they are very quick; one can begin to examine the phenotype of a knockdown organism generated using a MO (known as a morphant) within a few days of identifying a gene of interest. However, there are several disadvantages. MOs only transiently inhibit expression of a target gene. Also, they can be toxic, and cause non-specific phenotypes (302, 303). It can be challenging to confirm that expression of the target gene has been inhibited, and it can be even more challenging to confirm that the expression of no other off-target genes have been inhibited. There are methods to minimise the impact of these problems, so it is important to take them into account when designing experiments and interpreting results.

4.1.6 Animal models of dystroglycanopathy

The zebrafish orthologues of *LARGE*, *POMT1*, *POMT2*, *POMGnT1*, *FKTN*, and *FKRP* were identified in 2008 (304). All of these genes were expressed throughout early development, including in tissues relevant to dystroglycanopathy including the CNS, eye and muscle. Importantly, the authors also found that the antibody IIH6, which recognises glycosylated α-DG in humans, also works in zebrafish. Since then, in zebrafish embryos, dystroglycanopathy-associated genes have often been studied using MOs to knock down a candidate gene. For example, knockdown of *fkrp* results in changes to somites, muscle fibres, neuronal structures and eye shape, hypoglycosylation of α-DG and reduced laminin binding, all of which recapitulates the patients' phenotypes (17, 305). Similar results have been found for orthologues of many other dystroglycanopathy-associated genes including *POMT1*, *POMT2*, *B3GNT1*, *GTDC2*, *ISPD*, and *DAG1* (16, 254, 282, 306, 307).

As well as providing supportive evidence of causality, modelling a candidate dystroglycanopathy-associated gene in a model organism often leads to insights into the molecular pathology of disease that would not be possible on human patients. For example, examination of various tissues of the *Fktn* knockout mouse revealed that some of the structural tissue defects might be cause by disruption of the basal lamina, caused in turn by disruption of α-DG-ligand binding (308). In another important example, mass spectrometry of O-linked glycans in brain samples from mouse knockouts for *Pomgnt1*, *Large*, and *Dag1* yielded evidence that *Pomgnt1* may glycosylate other brain proteins in addition to α-DG for glycosylation. This finding could have important implications for understanding of pathology in patients (309). Experiments on zebrafish models of dystroglycanopathy have similarly revealed many insights into the disease pathology, such as FKTN and FKRP may have role in protein secretion (310).

Animal models can also lead to important developments in therapeutic strategies. For example, several groups have shown that overexpression of *Large* in mouse models can improve a dystroglycanopathy phenotype, even when the mouse has a pathogenic variant in another gene such as *Pomgnt1*, suggesting that overexpression of *LARGE* could be a viable therapeutic strategy for human dystroglycanopathy patients, whether or not their pathogenic variants are in *LARGE* (311, 312).

4.1.7 Aims, context, and colleagues

Some parts of this chapter have been published (313, 314). The parts of these two publications that I have reproduced in this chapter were all my work originally. This section briefly summarises the aspects of these two publications with which I was not directly involved, in order to put my own results into context.

The exomes of five patients with dystroglycanopathy were sequenced as part of the UK10K project, under the rare disease consortium, with the aim of identifying novel causative genes. Dr Sebahattin Cirak identified likely candidate genes for two of the patients. The genes were β-1,3-N-acetylgalactosaminyltransferase 2 (*B3GALNT2* [MIM 610194]) and guanosine diphosphate (GDP) mannose pyrophosphorylase B (*GMPPB* [MIM 615320]).

B3GALNT2 encodes a glycosyltransferase that is involved in the synthesis of GlcNac-β1,3GalNac, which is one of the glycans on α-DG, and is required for laminin binding (227, 315, 316). GMPPB catalyses the conversion of mannose-1-phosphate and GTP into GDP-mannose (317). GDP-mannose is required in four glycosylation pathways (Figure 4-2), including O-mannosylation of membrane and secretory glycoproteins, such as α-DG. Pathogenic variants in other members of this pathway can cause dystroglycanopathy (234, 235).

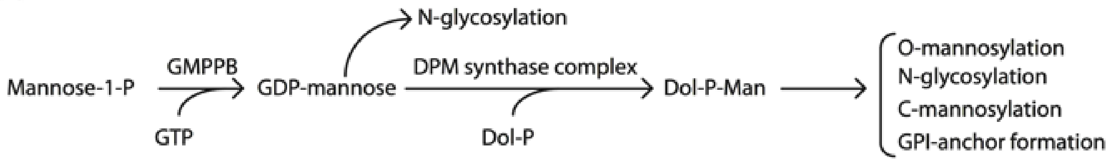

Figure 4 - 2: The function of GMPPB in glycosylation pathways.
This figure has been published in a modified form (313).

Through an international collaboration lead by Professor Francesco Muntoni and Dr A. Reghan Foley, six further individuals (two of whom were siblings) with dystroglycanopathy and *B3GALNT2* variants were identified, and seven further individuals with dystroglycanopathy and *GMPPB* variants were identified. All of the 15 patients had homozygous or compound heterozygous variants. The variants included missense, nonsense and frameshift changes.

All seven patients with *B3GALNT2* variants were severely affected. They had muscle dysfunction, cognitive impairment, and gross abnormalities of the brain with early presentation. Three of the patients also had epilepsy, and five had ophthalmologic abnormalities. Most reached no major motor milestones. In contrast, the phenotype of the eight patients with *GMPPB* variants was milder, and the presentation later. Most patients had hypotonia and poor muscle control, but they could walk. Most had intellectual delay, but mild, and structural brain abnormalities were only detected in three patients.

Dr Elizabeth Stevens and Dr Silvia Torelli confirmed hypoglycosylation of α-DG in most of the 15 patients using multiple methods, including immunoblot of both muscle protein lysate, and flow cytometry of fibroblasts, using the IIH6 antibody. They also transfected myoblasts with recombinant wildtype B3GALNT2, and showed that it localises to the ER, and that some of the variants identified in patients cause B3GALNT2 to mislocalise to the cytoplasm. In contrast, recombinant wildtype GMPPB localises to the cytoplasm, and some of the variants identified in patients result in formation of protein aggregates.

I carried out the work described in the rest of this chapter, under the overall supervision of Dr Derek Stemple and Dr Matthew Hurles, with direct, day-to-day supervision and assistance from Dr Yung-Yao Lin. Our aim was to knockdown orthologues of *B3GALNT2* and *GMPPB* in zebrafish embryos in order to recapitulate the patients' phenotypes to provide further evidence of pathogenicity, and to gain further insight into the pathology of the disease.

4.2 Materials and methods

4.2.1 Sequencing of clones

I obtained clones of *b3galnt2* (BC095777) and *gmppb* (BC078357.1) from Source BioScience (Nottingham, UK). I Sanger sequenced each insert using standard protocols, and confirmed that they did not contain any variants different from the human reference sequence. These sequences were used for all alignments and reagent design.

4.2.2 Reverse transcription polymerase chain reaction

To estimate the expression levels of zebrafish genes, I extracted RNA from 20–30 embryos with an RNeasy kit (QIAGEN, Crawley, UK), followed by reverse transcription with SuperScript III (Life Technologies). Polymerase chain reaction (PCR) was done with RedTaq DNA polymerase kit (Sigma-Aldrich, Dorset, UK). I amplified fragments of *b3galnt2* and *gmppb*, using *actb1* as a positive control. I visualised results using standard agarose gel electrophoresis. Primer sequences are listed in Table 4-1.

Gene	Primer function	Direction	Primer sequence (5'-3')
b3galnt2	Expression analysis	Forward	actcagagctccgcgatg
b3galnt2	Expression analysis	Reverse	cagagcagagatccctcaaa
gmppb	Expression analysis & MO-flanking	Forward	tacagcagcaggtcaatcgt
gmppb	Expression analysis	Reverse	acaacaatggtgccctctct
gmppb	MO-flanking	Reverse	gttctgcccaatcactgctg

Table 4 - 1: Primers used to analyse *b3galnt2* and *gmppb* expression in early zebrafish development, and splicing disruption in *gmppb* morphants.
The sequences of the *actb1* primers have been previously described (282). This table and legend have been published in a modified form (313).

4.2.3 Design and injection of morpholino oligonucleotides

For *b3galnt2*, I obtained one translation blocking (TB) MO and one splice blocking (SB) MO from Gene Tools, LLC (Philomath, OR, USA). For *gmppb*, I obtained one TB MO and three SB MOs. For MO sequences and the predicted effect of each SB MO, please see Table 4-2. All SB MOs were predicted to result in a frameshift. I also confirmed that there were no known variants in the MO binding site, and that the MO binding site was predicted to be specific.

Gene	**MO type**	**Sequence (5'-3')**	**Predicted Effect**
b3galnt2	TB	CGCCGCCGCTGCACTTCT**CAT**GGAC	NA
b3galnt2	SB	GGTCTGTctgtcaaggagaaataaa	Skipping of exon 2
gmppb	TB	CACCGACAAGAATCAGAGCTTT**CAT**	NA
gmppb	SB	gaaagactgccgtcagttacCTTGA	Retention of intron 2
gmppb	SB	GGACCAGctgaaaacagaaacagat	Skipping of exon 5
gmppb	SB	acagtgttcaaatcctttacCTTGC	Retention of intron 7

Table 4 - 2: Morpholino oligonucleotide sequences and predicted effects.
MO = morpholino oligonucleotide; TB = translation blocking; SB = splice blocking. In the sequences of TB MOs, the letters in bold indicate the start codon. In the sequences of SB MOs, the letters in lower case indicate the portion of the MO predicted to bind in the intron, and the letters in upper case indicate the portion of the MO predicted to bind in the exon. Sequences of *p53* and *dag1* MOs have been described (296, 302).

I injected MOs into 1- to 4- cell-stage Tuebingen Long Fin zebrafish embryos, which I reared as previously described (318). Unless otherwise stated, for *b3galnt2* I injected the TB MO at 4 ng along with a *p53* TB MO at 2 ng, and for *gmppb* I used the second SB MO, which is predicted to result in skipping of the fifth coding exon, at 3 ng dose, coinjected with *p53* MO 6 ng. To compare the eye diameter of different groups of embryos, I used unpaired two-tailed t tests.

4.2.4 Generation of green fluorescent protein-tagged RNA

I generated zebrafish expression plasmids (pCS2fl) containing cDNA of the gene of interest and a C-terminal green fluorescent protein (GFP) tag by Gateway Cloning Technology (Life Technologies, Paisley, UK) according to the manufacturer's instructions. I made mRNA for injection with mMessage mMachine SP6 kit (Ambion, for Life Technologies).

4.2.5 Immunofluorescence staining

I performed immunofluorescence staining on 48 hpf whole-mount embryos as previously described (296). For B3galnt2, I used primary antibodies against laminin (L-9393, Sigma-Aldrich) and β-DG (monoclonal, NCL-b-DG, Leica Microsystems, Milton Keynes, UK). For Gmppb I used primary antibodies against filamentous actin with the use of Alexa-Fluor-594-conjugated phalloidin (Life Technologies), β-DG, laminin, and IIH6.

4.2.6 Evans blue dye assay

For the Evans blue dye (EBD) assay, I immobilised 48 hpf live embryos in 1% low-melting-point agarose (Sigma-Aldrich) containing 0.016% tricaine (Sigma-Aldrich), and I injected a solution of 0.1% EBD (Sigma-Aldrich) into the pericardium. Two hours later, I examined them by confocal microscopy. To assess the significance of the results of the EBD assay, I used unpaired two-tailed t tests.

4.2.7 Immunoblotting

I performed microsome preparation and immunoblot analysis of zebrafish proteins as previously described (282), and quantified results using ImageJ software (319).

4.3 Results

4.3.1 *B3GALNT2* and *GMPPB* are conserved with their zebrafish orthologues

B3GALNT2 and *GMPPB* each have a single orthologue in the zebrafish. Zebrafish B3galnt2 (ENSDARP00000067823) is 53% identical to human B3GALNT2. Interestingly, the amino acid sequence of the galactosyltransferase domain alone is 68.5% identical between the two species (Table 4-3 and Figure 4-3). Zebrafish Gmppb (ENSDARP00000022618) is 81.4% identical to human GMPPB (Table 4-4 and Figure 4-4).

Species	Protein name	Ensembl identifier	% identity with human B3GALNT2 (whole protein)	% identity with human B3GALNT2 (galactosyl-transferase domain)
Pan troglodytes	B3GALNT2	ENSPTRT00000003896	99.6	100
Mus musculus	B3GALNT2	ENSMUST00000099747	88.4	93.3
Danio rerio	B3galnt2	Clone BC095777	54.6	68.5
Drosophila melanogaster	Beta-1,3-galactosyltransferase II	FBtr0088728	21.2	32.2
Caenorhabditis elegans	Beta-1,3-galactosyltransferase sqv-2	Y110A2AL_14_2	25.1	28.8

Table 4 - 3: Percentage identity of B3GALNT2 orthologues of five diverse eukaryotic species with human B3GALNT2.
Orthologues identified by BLAST alignment of the *Homo sapiens* B3GALNT2 sequence (ENSP00000355559) against the genomes of the species shown.

Species	Protein name	Ensembl identifier	% identity with human GMPPB
Pan troglodytes	GMPPB	ENSPTRP00000025773	99.7
Mus musculus	GMPPB	ENSMUSP00000107914	98.1
Danio rerio	Gmppb	ENSDARP00000022618	81.4
Drosophila melanogaster	CG1129	FBpp0078511	70.2
Caenorhabditis elegans	TAG-335	C42C1.5	63.8

Table 4 - 4: Percentage identity of GMPPB orthologues of five diverse eukaryotic species with human GMPPB.
Orthologues identified by BLAST alignment of the *Homo sapiens* GMPPB sequence (ENSP00000418565) against the genomes of the species shown. This table and legend have been published (313).

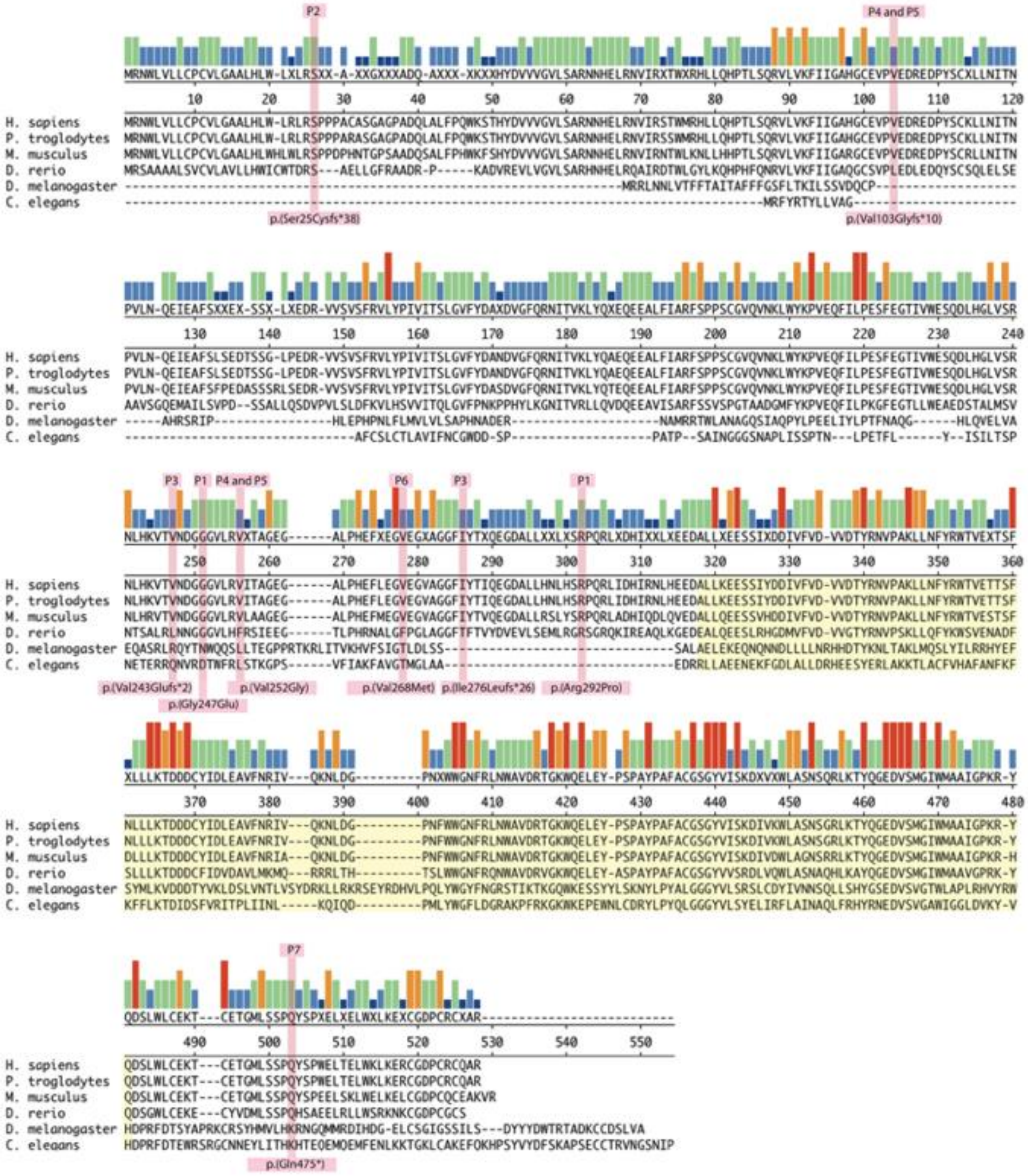

Figure 4 - 3: Protein alignment showing conservation of B3GALNT2.
Sequences are *Homo sapiens* (ENSP00000355559), *Pan troglodytes* (ENSPTRT00000003896), *Mus musculus* (ENSMUST00000099747), *Danio rerio* (clone BC095777), *Drosophila melanogaster* (FBtr0088728), and *Caenorhabditis elegans* (Y110A2AL_14_2). The height and colour of the bars indicates the degree of conservation of each amino acid residue, for example a red bar shows that a residue is conserved across all six species. The residues altered in the muscular dystrophy cases are highlighted in pink. The galactosyltransferase domain is highlighted in yellow.

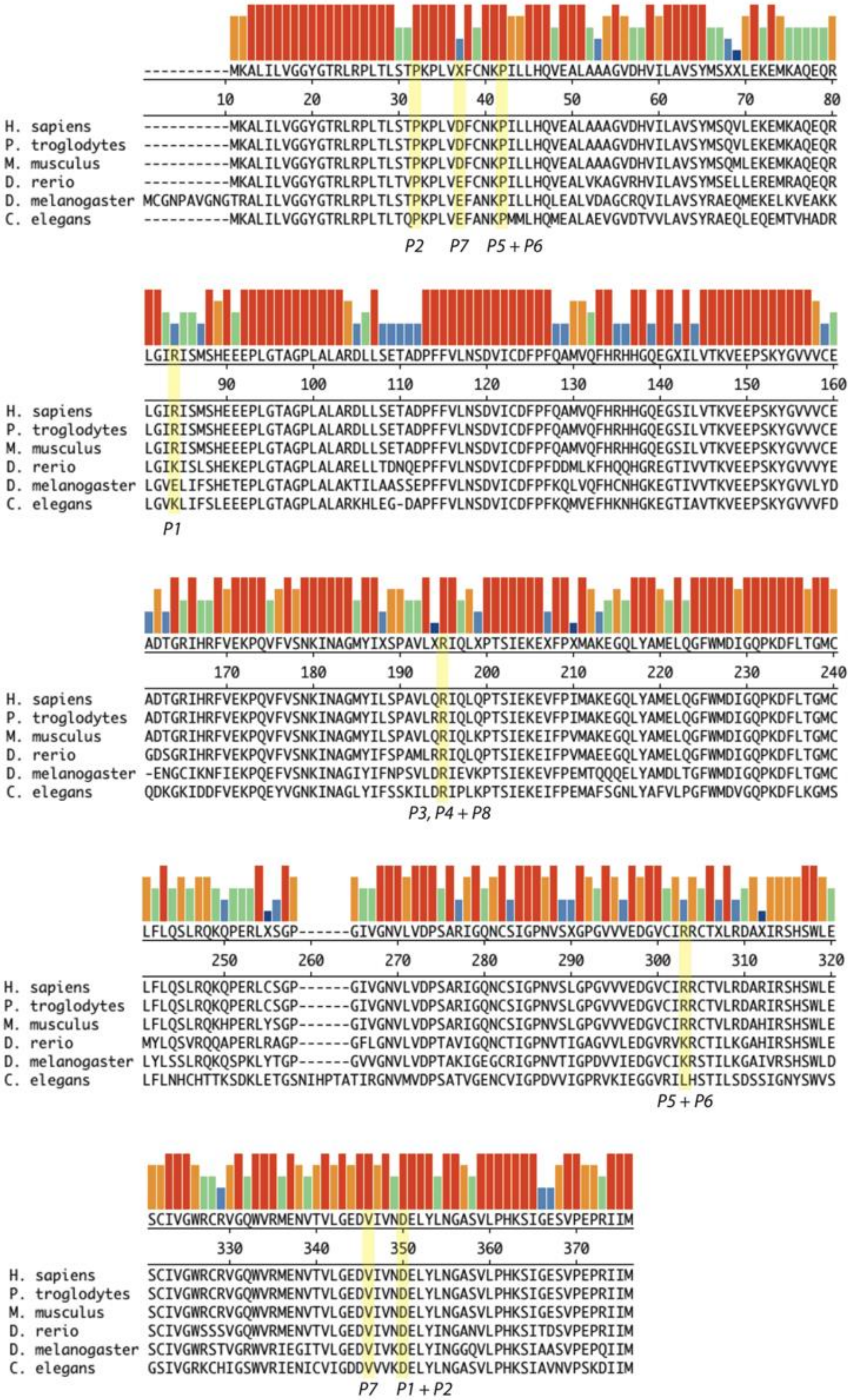
----------MKALILVGGYGTRLRPLTLSTPKPLVXFCNKPILLHQVEALAAAGVDHVILAVSYMSXXLEKEMKAQEQR
10 20 30 40 50 60 70 80
H. sapiens ----------MKALILVGGYGTRLRPLTLSTPKPLVDFCNKPILLHQVEALAAAGVDHVILAVSYMSQVLEKEMKAQEQR
P. troglodytes ----------MKALILVGGYGTRLRPLTLSTPKPLVDFCNKPILLHQVEALAAAGVDHVILAVSYMSQVLEKEMKAQEQR
M. musculus ----------MKALILVGGYGTRLRPLTLSTPKPLVDFCNKPILLHQVEALAAAGVDHVILAVSYMSQMLEKEMKAQEQR
D. rerio ----------MKALILVGGYGTRLRPLTLTVPKPLVEFCNKPILLHQVEALVKAGVRHVILAVSYMSELLEREMRAQEQR
D. melanogaster MCGNPAVGNGTRALILVGGYGTRLRPLTLSTPKPLVEFANKPILLHQLEALVDAGCRQVILAVSYRAEQMEKELKVEAKK
C. elegans ----------MKALILVGGYGTRLRPLTLTQPKPLVEFANKPMMLHQMEALAEVGVDTVVLAVSYRAEQLEQEMTVHADR
P2 P7 P5 + P6
LGIRISMSHEEEPLGTAGPLALARDLLSETADPFFVLNSDVICDFPFQAMVQFHRHHGQEGXILVTKVEEPSKYGVVVCE
90 100 110 120 130 140 150 160
H. sapiens LGIRISMSHEEEPLGTAGPLALARDLLSETADPFFVLNSDVICDFPFQAMVQFHRHHGQEGSILVTKVEEPSKYGVVVCE
P. troglodytes LGIRISMSHEEEPLGTAGPLALARDLLSETADPFFVLNSDVICDFPFQAMVQFHRHHGQEGSILVTKVEEPSKYGVVVCE
M. musculus LGIRISMSHEEEPLGTAGPLALARDLLSETADPFFVLNSDVICDFPFQAMVQFHRHHGQEGSILVTKVEEPSKYGVVVCE
D. rerio LGIKISLSHEKEPLGTAGPLALARELLTDNQEPFFVLNSDVICDFPFDDMLKFHQQHGREGTIVVTKVEEPSKYGVVVYE
D. melanogaster LGVELIFSHETEPLGTAGPLALAKTILAASSEPFFVLNSDVICDFPFKQLVQFHCNHGKEGTIVVTKVEEPSKYGVVLYD
C. elegans LGVKLIFSLEEEPLGTAGPLALARKHLEG-DAPFFVLNSDVICDFPFKQMVEFHKNHGKEGTIAVTKVEEPSKYGVVVFD
P1
ADTGRIHRFVEKPQVFVSNKINAGMYIXSPAVLXRIQLXPTSIEKEXFPXMAKEGQLYAMELQGFWMDIGQPKDFLTGMC
170 180 190 200 210 220 230 240
H. sapiens ADTGRIHRFVEKPQVFVSNKINAGMYILSPAVLQRIQLQPTSIEKEVFPIMAKEGQLYAMELQGFWMDIGQPKDFLTGMC
P. troglodytes ADTGRIHRFVEKPQVFVSNKINAGMYILSPAVLRRIQLQPTSIEKEVFPIMAKEGQLYAMELQGFWMDIGQPKDFLTGMC
M. musculus ADTGRIHRFVEKPQVFVSNKINAGMYILSPAVLQRIQLKPTSIEKEIFPVMAKEGQLYAMELQGFWMDIGQPKDFLTGMC
D. rerio GDSGRIHRFVEKPQVFVSNKINAGMYIFSPAMLRRIQLQPTSIEKEIFPVMAEEGQLYAMELQGFWMDIGQPKDFLTGMC
D. melanogaster -ENGCIKNFIEKPQEFVSNKINAGIYIFNPSVLDRIEVKPTSIEKEVFPEMTQQQELYAMDLTGFWMDIGQPKDFLTGMC
C. elegans QDKGKIDDFVEKPQEYVGNKINAGLYIFSSKILDRIPLKPTSIEKEIFPEMAFSGNLYAFVLPGFWMDVGQPKDFLKGMS
P3, P4 + P8
LFLQSLRQKQPERLXSGP------GIVGNVLVDPSARIGQNCSIGPNVSXGPGVVVEDGVCIRRCTXLRDAXIRSHSWLE
250 260 270 280 290 300 310 320
H. sapiens LFLQSLRQKQPERLCSGP------GIVGNVLVDPSARIGQNCSIGPNVSLGPGVVVEDGVCIRRCTVLRDARIRSHSWLE
P. troglodytes LFLQSLRQKQPERLCSGP------GIVGNVLVDPSARIGQNCSIGPNVSLGPGVVVEDGVCIRRCTVLRDARIRSHSWLE
M. musculus LFLQSLRQKHPERLYSGP------GIVGNVLVDPSARIGQNCSIGPNVSLGPGVVVEDGVCIRRCTVLRDAHIRSHSWLE
D. rerio MYLQSVRQQAPERLRAGP------GFLGNVLVDPTAVIGQNCTIGPNVTIGAGVVLEDGVRVKRCTILKGAHIRSHSWLE
D. melanogaster LYLSSLRQKQSPKLYTGP------GVVGNVLVDPTAKIGEGCRIGPNVTIGPDVVIEDGVCIKRSTILKGAIVRSHSWLD
C. elegans LFLNHCHTTKSDKLETGSNIHPTATIRGNVMVDPSATVGENCVIGPDVVIGPRVKIEGGVRILHSTILSDSSIGNYSWVS
P5 + P6
SCIVGWRCRVGQWVRMENVTVLGEDVIVNDELYLNGASVLPHKSIGESVPEPRIIM
330 340 350 360 370
H. sapiens SCIVGWRCRVGQWVRMENVTVLGEDVIVNDELYLNGASVLPHKSIGESVPEPRIIM
P. troglodytes SCIVGWRCRVGQWVRMENVTVLGEDVIVNDELYLNGASVLPHKSIGESVPEPRIIM
M. musculus SCIVGWRCRVGQWVRMENVTVLGEDVIVNDELYLNGASVLPHKSIGESVPEPRIIM
D. rerio SCIVGWSSSVGQWVRMENVTVLGEDVIVNDELYINGANVLPHKSITDSVPEPRIIM
D. melanogaster SCIVGWRSTVGRWVRIEGITVLGEDVIVKDELYINGGQVLPHKSIAASVPEPQIIM
C. elegans GSIVGRKCHIGSWVRIENICVIGDDVVVKDELYLNGASVLPHKSIAVNVPSKDIIM
P7 P1 + P2

Figure 4 - 4: Protein alignment showing conservation of GMPPB.
Sequences are *Homo sapiens* (ENSP00000418565), *Pan troglodytes* (ENSPTRP00000025773), *Mus musculus* (ENSMUSP00000107914), *Danio rerio* (ENSDARP00000022618), *Drosophila melanogaster* (FBpp0078511), and *Caenorhabditis elegans* (C42C1.5). The height and colour of the bars indicates the degree of conservation of each amino acid residue. The residues altered in the muscular dystrophy cases are highlighted in yellow. This figure and legend have been published (313).

4.3.2 Expression of *b3galnt2* and *gmppb* throughout early zebrafish development

To investigate temporal expression patterns in zebrafish development I reverse transcribed RNA extracted from zebrafish embryos at various stages of development, and amplified a fragment of each gene of interest. Both *b3galnt2* and *gmppb* are clearly expressed at each of the stages tested (Figure 4-5).

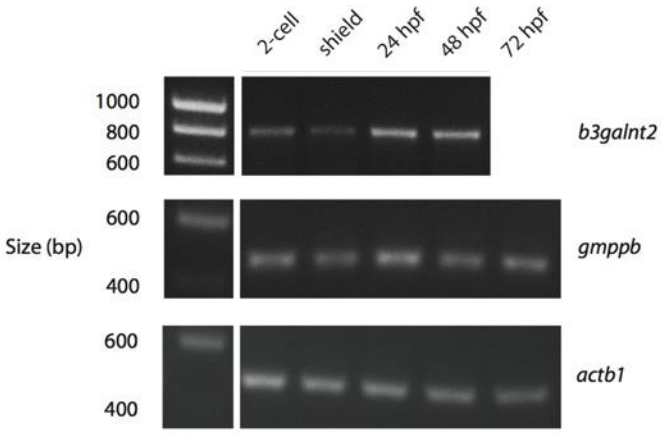

Figure 4 - 5: Reverse transcription PCR shows expression of *b3galnt2* and *gmppb* throughout early zebrafish development.
I performed reverse transcription PCR (RT-PCR) on RNA extracted from wildtype zebrafish embryos at the five developmental stages indicated. Expression of *b3galnt2* and *gmppb* was detected at each stage, using PCR amplification of cDNA fragments. An *actb1* fragment indicates near-equivalent amounts of input cDNA. This figure and legend have been published in a modified form (313, 314).

This is consistent with published *in-situ* hybridisation data suggesting that *b3galnt2* is ubiquitously expressed at early stages and becomes more anteriorly localised by 60 hpf (320), and that at early developmental stages (1-4 somites to 10-13 somites) *gmppb* is expressed primarily in the notochord, periderm and polster, but by prim-15 stage, expression is primarily localised to the head (321).

4.3.3 Finding the optimal morpholino oligonucleotide and dose for *b3galnt2* and *gmppb*

I injected a range of quantities of each MO into zebrafish embryos, and examined their phenotype, to determine the optimal MO and dose for each gene of interest (data not shown). For *b3galnt2* I found that both MOs resulted in similar phenotypes, suggesting they do specifically knock down *b3galnt2*. However, even at very low doses (2.5 ng) the SB MO had such a severe effect on muscle development that normal structures including fibres and myosepta were not discernable, whereas the TB MO had a clearer and more moderate phenotype.

A proportion of MOs cause non-specific *p53* upregulation. This results in phenotypes that can be mistakenly attributed to knockdown of the gene of interest (302, 303). I tested whether this would occur with the *b3galnt2* TB MO, by injecting 4 ng of the TB MO with or without a *p53* TB MO at 2 ng, and comparing the resulting phenotypes. At 24 hpf there was some neurodegeneration in the brains of the embryos without the *p53* MO, which was not present where the *p53* MO had been coinjected (Figure 4-6). Therefore, unless otherwise stated, for *b3galnt2* the TB MO was injected at 4 ng along with a *p53* TB MO at 2 ng.

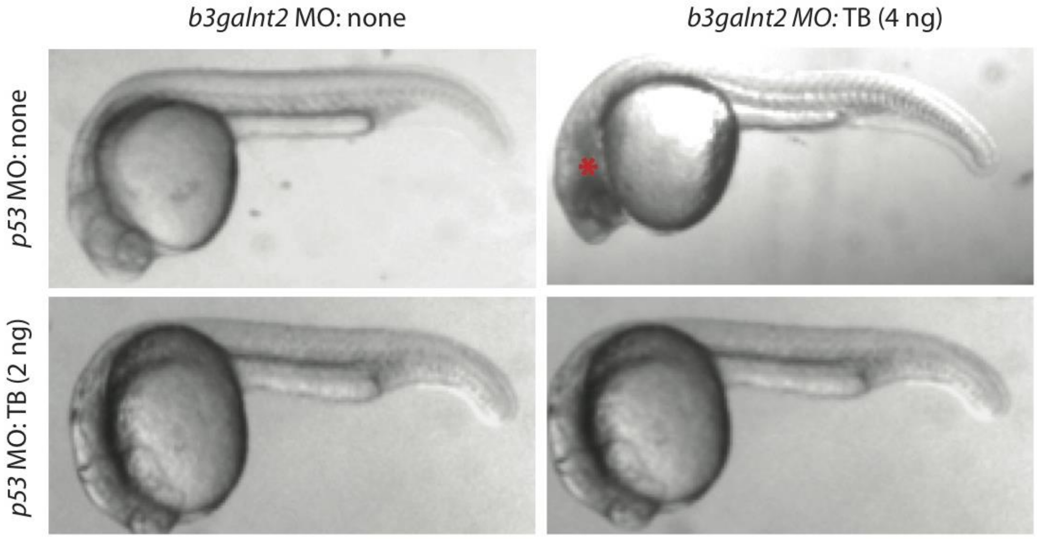

Figure 4 - 6: Coinjection of *p53* MO rescues the neurodegeneration induced by *b3galnt2* MO.
Injection of the *b3galnt2* TB MO (4 ng) on its own induces neurodegeneration, which makes the brain appear cloudy and dark (red asterisk). This is rescued then *p53* TB MO (2 ng) is coinjected, indicating that neurodegeneration is a non-specific phenotype (302, 303).

For *gmppb*, the TB MO and the first SB MO gave no discernable phenotype, even at very high doses (12.5 ng). The third SB MO also gave no phenotype. The second SB MO, which is predicted to result in skipping of the fifth coding exon**,** gave a consistent and relevant phenotype at reasonable doses, without excessive lethality. Therefore, unless otherwise stated, the second SB MO was used at 3 ng dose, coinjected with *p53* MO 6 ng.

4.3.4 Morpholino oligonucleotides reduce the expression of *b3galnt2* and *gmppb*

To confirm the efficacy and specificity of the *b3galnt2* MO I cloned *b3galnt2* into a GFP expression vector, made RNA from this, and injected it into zebrafish embryos (25 pg). At 24 hpf I showed using confocal microscopy that these embryos express the wildtype recombinant GFP-tagged *b3galnt2* RNA. This was suppressed when coinjected with *b3galnt2* MO (Figure 4-7), indicating that the *b3galnt2* MO effectively knocks down *b3galnt2*.

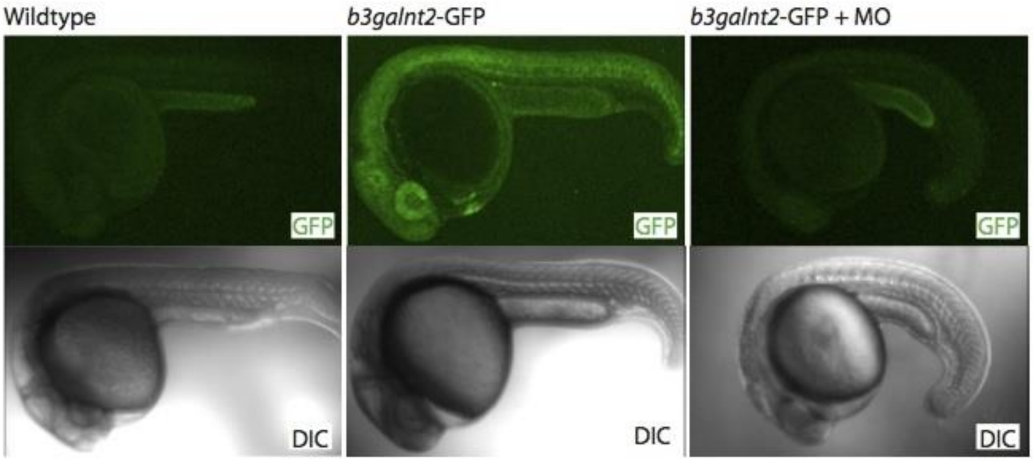

Figure 4 - 7: The *b3galnt2* MO inhibits expression of recombinant GFP-tagged *b3galnt2* RNA.
Embryos express wildtype recombinant GFP-tagged *b3galnt2* RNA (25 pg). This is suppressed when coinjected with the MO (*b3galnt2* TB 4 ng coinjected with *p53* TB 2 ng). Photographs were taken at 24 hpf. This figure and legend have been published (314).

To test the efficacy of the *gmppb* MO, I extracted RNA from wildtype and MO-injected embryos and performed reverse transcription PCR (RT-PCR) with primers flanking the MO binding site. I found that the morphants had a lower abundance of gmppb cDNA than wildtype, whereas the abundance of cDNA of a housekeeping gene (*actb1*) was approximately equal in the two groups, indicating that the *gmppb* MO effectively knocks down *gmppb* (Figure 4-8).

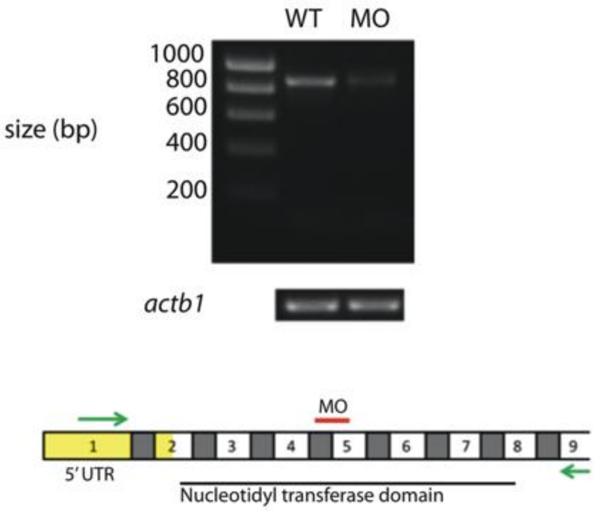

Figure 4 - 8: The *gmppb* splice blocking MO disrupts RNA splicing.
I extracted RNA from 48 hpf wildtype zebrafish embryos (WT) alongside embryos injected with the *gmppb* SB MO (2.5 ng, 5 ng and 7.5 ng pooled). I performed RT-PCR and PCR amplified a ~900 bp fragment of *gmppb* cDNA using primers (indicated by green arrows on schematic diagram) that bind either side of the MO binding site (indicated by red line). A clear reduction in band intensity is seen in the MO embryos, indicating that the MO disrupts correct mRNA splicing. An *actb1* fragment indicates near-equivalent amounts of input cDNA. This figure and legend have been published (313).

4.3.5 *b3galnt2* morphants have gross morphological defects including hydrocephalus and impaired motility

I examined the gross morphological phenotype of *b3galnt2* knockdown embryos at 48 hpf using light microscopy, and compared it to wildtype embryos. Consistently observed phenotypes were hydrocephalus, curvature of the tail, severely impaired motility, mild retinal degeneration, growth restriction, pericardial effusion, enlarged and dysmorphic yolk, delayed or failed hatching, and mild hypopigmentation (Figure 4-9A).

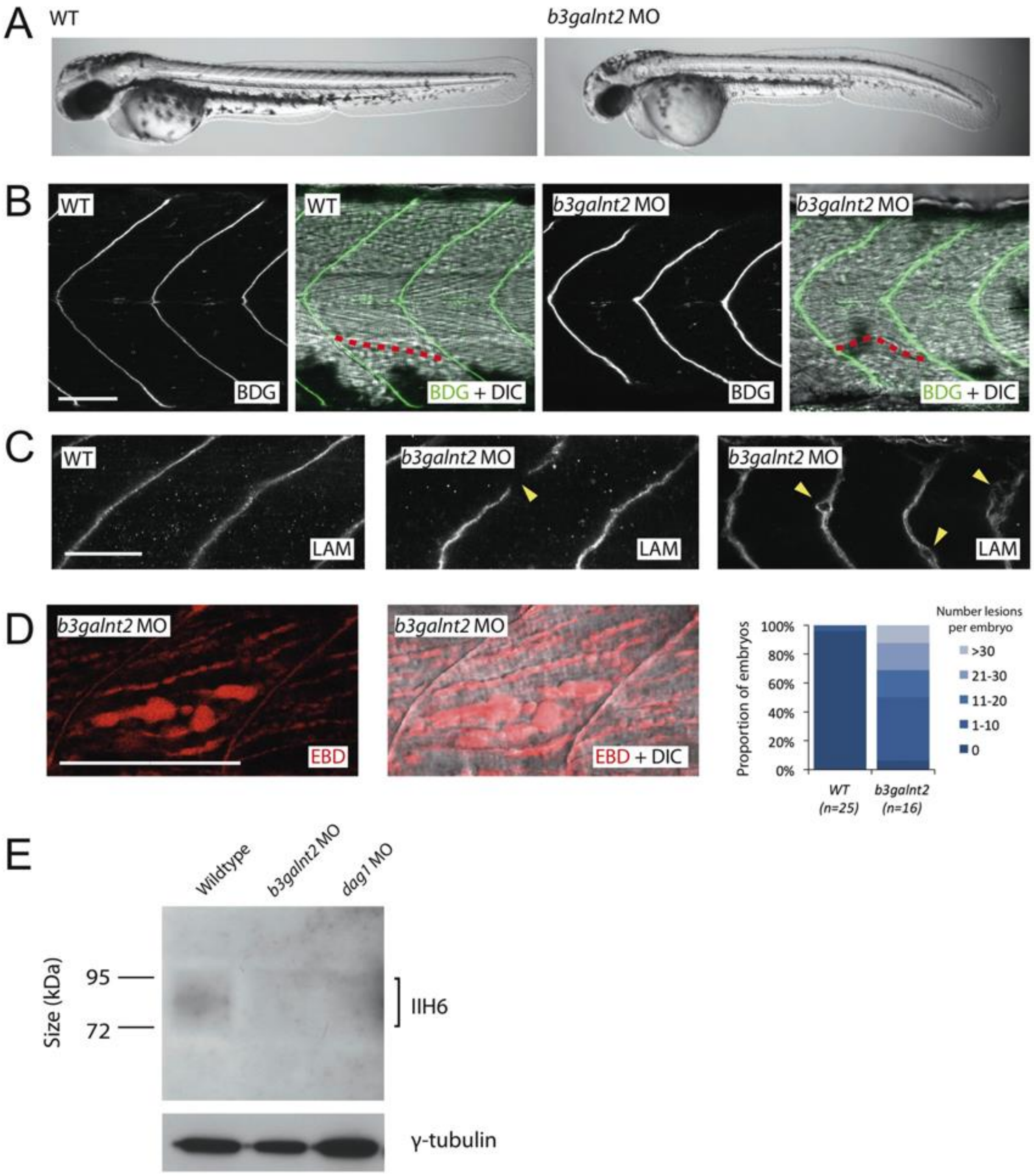

Figure 4 - 9: *b3galnt2* knockdown zebrafish embryos have muscle defects and hypoglycosylated α-DG at 48 hpf.

(A) Whole-mount pictures of live embryos show gross morphological defects. **(B)** Immunofluorescence staining by an antibody against β-DG and differential interference contrast (DIC) microscopy showed that the muscle fibres are disordered. One sample fibre is highlighted in red. **(C)** Immunofluorescence staining by an antibody against laminin (LAM) shows gaps in the myosepta of knockdown embryos and degeneration of the ECM. **(D)** Evans blue dye assay (EBD) highlights frequent lesions between muscle fibres in *b3galnt2* morphants, which are very rarely seen in wildtype embryos. **(A–D)** *b3galnt2* MO: *b3galnt2* TB MO (4 ng) coinjected with *p53* TB MO (2 ng). Scale bars represent 50 mm. **(E)** Immunoblotting by isolated microsome protein from 48 hpf embryos and IIH6 antibody showed a reduction in glycosylated α-DG in the *b3galnt2* knockdown embryos. *b3galnt2* MO: *b3galnt2* TB MO (5 ng) coinjected with *p53* TB MO (2.5 ng). *dag1* MO: *dag1* TB MO (5 ng). This figure and legend have been published (314).

The occurrence of hydrocephalus is particularly interesting as hydrocephalus was observed in three of the muscular dystrophy patients. Hydrocephalus is also caused by knocking down other dystroglycanopathy-associated genes such as *fktn*, *fkrp*, and *ispd* in zebrafish embryos (282). I counted the number of embryos that had hydrocephalus in a group of morphants and a group of wildtype embryos. 72% of *b3galnt2* morphants had some degree of hydrocephalus compared to 0% of wild type (Table 4-5). This is a highly significant difference (p = 2.2 x 10^{-16}, Fisher's exact test).

	Wildtype	**Morphant**
Without hydrocephalus	72 (100%)	11 (28%)
With hydrocephalus	0 (0%)	28 (72%)
Total	72	39

Table 4 - 5: b3galnt2 morphants are significantly more likely to have hydrocephalus than wildtype embryos.
Morphant = *b3galnt2* TB 4ng + *p53* 2ng. Classified according to appearance under light microscope at 48 hpf. Fisher exact test; p = 2.2 x 10^{-16}.

4.3.6 *b3galnt2* morphants have muscle defects including gaps in the myosepta and lesions between fibres

To characterise the muscle phenotype, I performed immunofluorescence staining. Compared to the chevron-shaped somite boundaries flanking straight muscle fibres in wildtype embryos, *b3galnt2* morphants consistently showed slightly U-shaped somites and disordered muscle fibres (Figure 4-9B). Laminin staining revealed occasional gaps in the myosepta (the connective tissues where the muscle fibres anchor), suggesting disruption of the ECM (Figure 4-9C).

Next, I quantified the severity of muscle damage by the EBD assay. EBD is an azo dye that binds to proteins such as albumin and is transported in the serum. It fluoresces upon protein binding and infiltrates muscle, where it penetrates compromised sarcolemma and accumulates at lesions between muscle fibres (interfibre spaces) (310, 322). Compared to wildtype embryos, *b3galnt2* morphants showed more severely damaged muscle, with increased number of lesions, ranging from less than 10 to more than 30 lesions per embryo (Figure 4-9D).

4.3.7 *b3galnt2* morphants have hypoglycosylated α-dystroglycan

Immunoblot analysis with the IIH6 antibody on protein extracts from wildtype embryos and *b3galnt2* morphants showed a reduction of the IIH6 signal in *b3galnt2* morphants, indicating that knockdown of *b3galnt2* led to reduced functional glycosylation of α-DG (Figure 4-9E). This is consistent with the human data and strongly suggests that this may be the molecular mechanism behind the phenotypes described.

4.3.8 Coinjection with wildtype RNA fails to rescue the *b3galnt2* morphant phenotype

In experiments using MOs, coinjection of wildtype RNA of the targeted gene can reduce the severity of the phenotype, providing supporting evidence that the phenotype is specific to knockdown of the gene of interest (323). To this end, I injected GFP-tagged wildtype human *B3GALNT2* mRNA (100 pg) into the cell of zebrafish embryos at the 1-cell stage, along with the MO. At 24 hpf I confirmed the expression of wildtype recombinant GFP-tagged *B3GALNT2* RNA by fluorescence microscopy. At 48 hpf I phenotyped the embryos by examining gross morphology, performing the EBD assay, and measuring the diameter of the eyes.

Human wildtype *B3GALNT2* RNA did not rescue the phenotype of *b3galnt2* zebrafish morphants. Compared to controls that had only been injected with *b3galnt2* MO, embryos that had been injected with *b3galnt2* MO and GFP-tagged *B3GALNT2* RNA had slightly more severe morphological defects such as curvature (Figure 4-10A). They had severe muscle lesions, with a slightly higher median number of lesions per embryo than controls that had only been injected with *b3galnt2* MO (Figure 4-10B). This difference is not statistically significant (p=0.2). They had a reduction in eye diameter compared to wildtype embryos and controls that had only been injected with GFP-tagged *B3GALNT2* RNA, which was not significantly different to that seen in controls that had only been injected with *b3galnt2* MO (p=0.26) (Figure 4-10C). Finally, a slightly higher proportion of embryos that had been injected with *b3galnt2* MO and GFP-tagged *B3GALNT2* RNA had hydrocephalus than controls that had only been injected with *b3galnt2* MO (Figure 4-10D). Control embryos that had only been injected with GFP-tagged *B3GALNT2* RNA were indistinguishable from wildtype embryos, except that a small proportion of them had mild hydrocephalus, whereas none of the uninjected embryos did (Figure 4-10D). Increasing the injected dose of RNA to 200 pg also did not result in phenotypic rescue (data not shown).

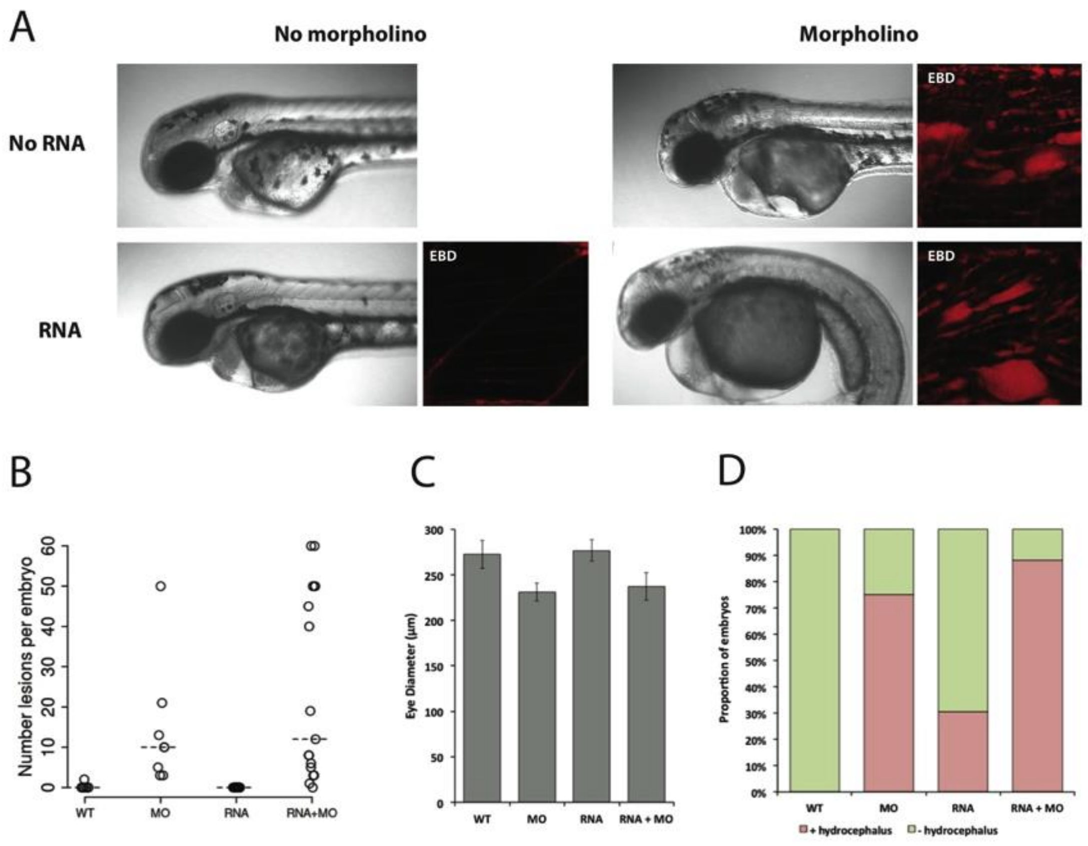

Figure 4 - 10: Coinjection with wildtype human *B3GALNT2* RNA fails to rescue the *b3galnt2* morphant phenotype.
48 hpf zebrafish embryos. Expression of GFP was confirmed at 24 hpf. **(A-D)** MO or Morpholino = *b3galnt2* TB MO (4 ng) coinjected with *p53* TB MO (2 ng). RNA = GFP-tagged wildtype human *B3GALNT2* mRNA (100 pg), WT = wildtype. Sample size for the WT, MO, RNA and RNA+MO groups respectively are 9, 8, 23 and 17 embryos. **(A)** Phenotype of anterior portion of embryos photographed using a confocal microscope. EBD = Evans blue dye assay. The appearance of pericardial oedema may be exaggerated in these images due to the injection of EBD. No photograph of the EBD assay on wildtype embryos was taken in this experiment. **(B)** Number of lesions. Dashed horizontal line indicates median. **(C)** Diameter of eyes. **(D)** Proportion of embryos that have hydrocephalus.

I tested whether the divergence between human and zebrafish orthologues of B3GALNT2 could explain this failure to rescue, by injecting GFP-tagged wildtype zebrafish *b3galnt2* mRNA (200 pg) into zebrafish embryos, along with the *b3galnt2* MO, and performing the EBD assay at 48 hpf. GFP-tagged wildtype zebrafish *b3galnt2* mRNA resulted in no improvement in the gross morphology of the embryos, or the appearance of muscle lesions (Figure 4-11).

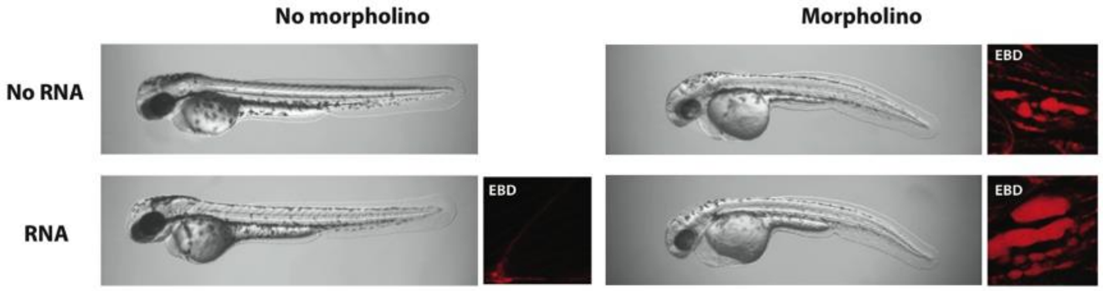

Figure 4 - 11: Coinjection with wildtype zebrafish *b3galnt2* RNA fails to rescue the *b3galnt2* morphant phenotype.
48 hpf zebrafish embryos. Expression of GFP was confirmed at 24 hpf. Morpholino = *b3galnt2* TB MO (4 ng) coinjected with *p53* TB MO (2 ng). RNA = GFP-tagged wildtype zebrafish *b3galnt2* mRNA (200 pg). Phenotype of embryos photographed using light microscope. EBD=Evans blue dye assay. EBD assay was not performed on wildtype embryos in this experiment. Each whole embryo picture is representative of at least 20 embryos, and each EBD picture is representative of at least five embryos.

4.3.9 *gmppb* morphants have gross morphological defects including micropthalmia and impaired motility

I examined the gross morphological phenotype of *gmppb* knockdown embryos at 48 hpf using light microscopy, and compared it to wildtype embryos. Morphologically, *gmppb* morphants were shorter than wildtype uninjected embryos at 48 hpf and often had bent tails. Other phenotypes included hypopigmentation, micropthalmia, hydrocephalus, increased lethality, and reduced motility (Figure 4-12A). The difference in diameter of the eyes of wildtype and *gmppb* morphant embryos was statistically significant, following correction for body length ($p < 1 \times 10^{-7}$; Figure 4-12B). Although none of the dystroglycanopathy cases reported micropthalmia, this is a phenotype that is common in individuals with severe forms of CMD, such as WWS and MEB (216, 221, 254, 281, 282, 314).

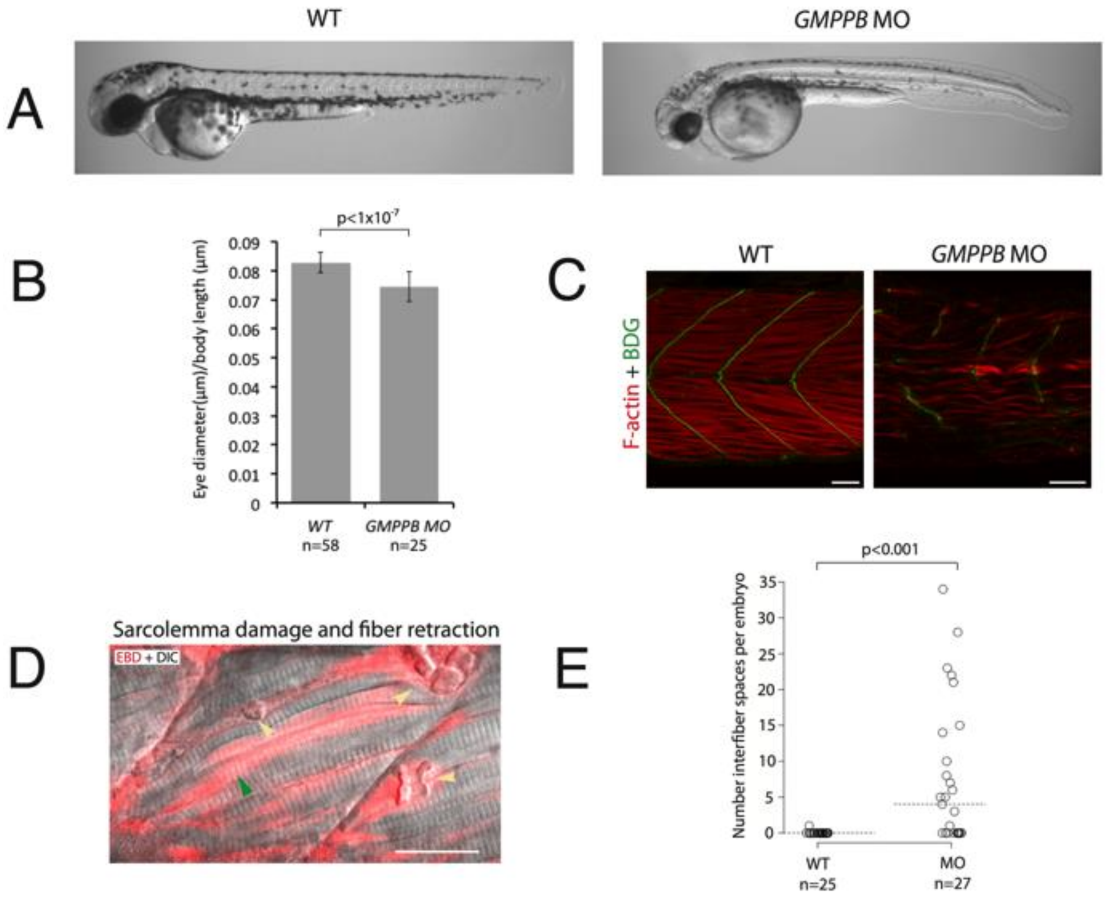

Figure 4 - 12: *gmppb* knockdown zebrafish embryos have morphological defects, damaged muscle, and hypoglycosylated α-DG at 48 hpf.

(A) Bright-field microscopy of live embryos shows morphological defects of the *gmppb*-MO-injected embryos (injected with *gmppb* splice-blocking MO 3 ng + p53 MO 6 ng) as compared to uninjected wildtype embryos. **(B)** I measured the eye diameter of *gmppb* morphants and wildtype embryos, normalised each eye diameter measurement to the embryo's body length, and assessed statistical significance using an unpaired two-tailed t-test. **(C)** Phalloidin staining of filamentous actin (red) and immunostaining with an antibody against β-DG (green). **(D)** Live *gmppb*-MO-injected embryos injected with EBD (red) and imaged by confocal microscopy. Some fibres showed EBD infiltration, indicating damage to the sarcolemma (green arrow), and other fibres detached from the myosepta and retracted (yellow arrow) and thus left a space. The following abbreviation is used: DIC, differential interference contrast. The scale bar represents 25 mm. **(E)** *gmppb* morphants have significantly more interfibre spaces than do wildtype uninjected embryos. The horizontal dotted line shows the median. This figure and legend have been published in a modified form (313).

4.3.10 *gmppb* morphants have muscle defects including disordered fibres, incomplete myosepta, and interfibre spaces

To characterise muscle defects in *gmppb* knockdown zebrafish embryos, I used phalloidin to label filamentous actin, along with immunostaining with antibodies against β-DG (which localises to the myosepta, the connective tissue to which muscle fibres anchor). I observed that the muscle fibres in *gmppb* morphants were sparse and disordered. Furthermore, fibres were frequently observed to span two somites, indicating damage or incomplete development of the myosepta (Figure 4-12C).

To further explore the muscle phenotypes in *gmppb* morphants, I injected EBD into the pericardium of embryos at 48 hpf. Compared with uninjected control embryos, *gmppb* morphants had significantly more EBD accumulation within interfibre spaces ($p < 0.001$; Figure 4-12D-E). In addition, EBD infiltrated both retracted and some intact muscle fibres in *gmppb*-knockdown embryos, suggesting that sarcolemma integrity was compromised prior to muscle fibre breakdown.

4.3.11 *gmppb* morphants have hypoglycosylated α-dystroglycan

Next, I investigated whether the laminin-binding glycan on α-DG is reduced in *gmppb* morphants. To do this, I performed immunoblots with the IIH6 antibody on membrane proteins enriched from wildtype embryos and *gmppb* morphants, as well as *dag1* morphants as a negative control. After normalisation to γ-tubulin loading control, *gmppb* morphants showed a slight but clear reduction in IIH6 levels (71% of that of the wildtype embryos) and *dag1* morphants showed a strong reduction in IIH6 (15% of that of the wild-type embryos) (Figure 4-13A). To confirm this finding, I performed double immunostaining with the IIH6 antibody and an antibody against laminins. In wildtype embryos, laminin and glycosylated α-DG colocalised at the myosepta. In *gmppb* morphants, the IIH6 staining was severely reduced, and laminin staining revealed widened myosepta, indicating a reduction in glycosylation of a-DG associated with abnormal basement-membrane structure (Figure 4-13B).

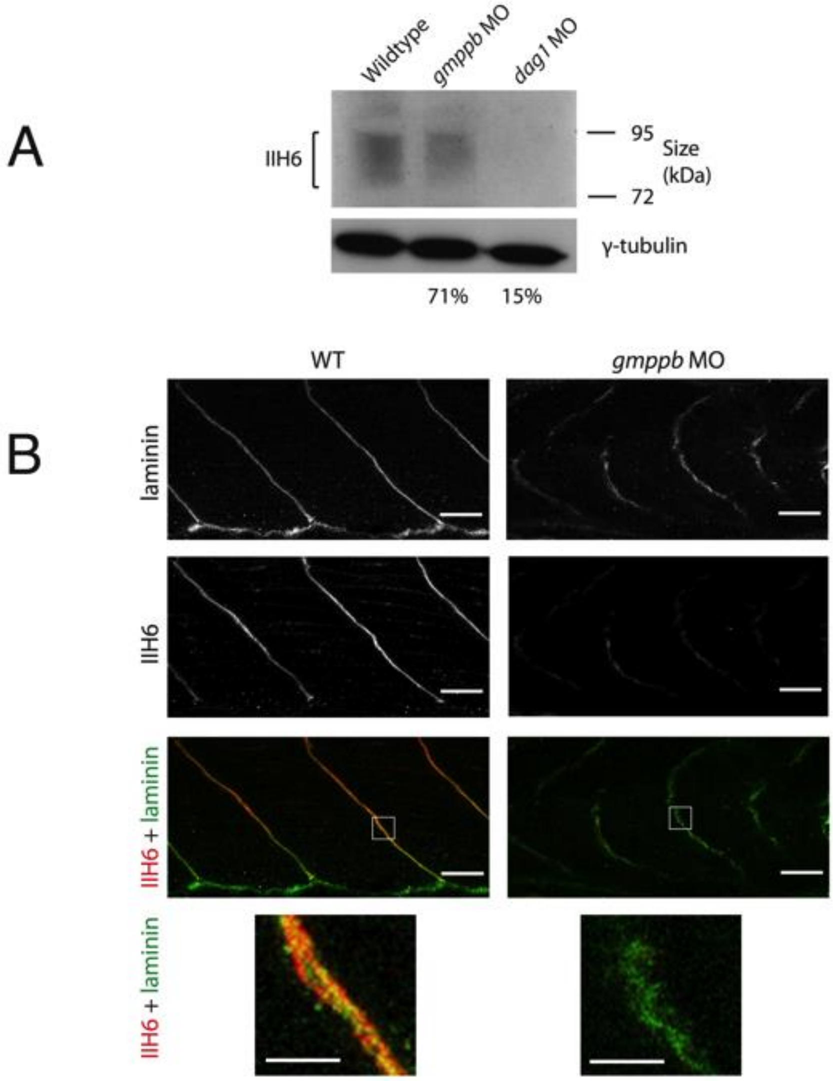

Figure 4 - 13: α-DG of *gmppb* morphants is hypoglycosylated relative to wildtype embryos.

(A) An immunoblot shows a reduction in α-DG glycosylation. Percentage figures indicate the intensity of morphant bands relative to that of the wildtype and are adjusted for the γ-tubulin loading control. ''*gmppb* MO'' indicates embryos injected with *gmppb* MO 3 ng + p53 MO 6 ng, and ''*dag1* MO'' indicates embryos injected with *dag1* TB MO (5 ng). **(B)** Antibodies used are against laminins, (polyclonal antibody L9393, Sigma-Aldrich, Dorset, UK), and glycosylated α-DG (monoclonal antibody IIH6, from Professor Kevin Campbell). Staining of laminins revealed abnormal myosepta structure, while the fluorescent intensity of IIH6 epitope was significantly reduced in *gmppb* morphants. White square indicates magnified area. Scale bar = 25 μm except for magnified area where scale bar = 6 μm. This figure and legend have been published in a modified form (313).

4.4 Discussion

4.4.1 Summary

Zebrafish are an appropriate model with which to study many human diseases, and have been particularly fruitful in the study of muscular dystrophy such as dystroglycanopathy. Here I have shown that zebrafish are an appropriate model with which to study the two dystroglycanopathy-associated genes *B3GALNT2* and *GMPPB*, because of moderate or high levels of conservation respectively, and appropriate temporal expression. I designed MOs to knock down the orthologues of both genes, and I optimised and validated the MOs. I performed extensive characterisation of the phenotypes of my two models. In each case, I demonstrate that aspects of the patients' phenotypes have been recapitulated in the zebrafish embryos at three levels: gross morphology, muscle structure and molecular level (hypoglycosylation of α-DG). This contributes to the body of data, which also includes the clinical and cellular data generated by colleagues as described in section 4.1.7, that supports the conclusion that pathogenic variants in *B3GALNT2* or *GMPPB* can cause dystroglycanopathy.

4.4.2 Phenotypic rescue

One method of confirming that knockdown of a gene of interest specifically causes a morphant phenotype, is to coinject wildtype RNA of the targeted gene and demonstrate reduction in the severity of the phenotype (323). The RNA can be from the zebrafish gene, or, if there is sufficient homology between zebrafish and humans, RNA from the human orthologue can be used. This letter method has the advantage that, if rescue is achieved, one can next demonstrate failure to rescue with a construct containing the patient's variant, providing compelling evidence of pathogenicity of that variant (254, 324).

In this study, I have demonstrated that neither zebrafish *b3galnt2* RNA nor human *B3GALNT2* RNA rescues the phenotype of *b3galnt2* morphant zebrafish embryos. This failure to rescue in no way suggests that the phenotypes I have detailed are *not* specific to knockdown of *b3galnt2*. While expression of endogenous genes is subject to exquisitely specific and complex spatial-temporal regulation, rescue experiments induce ubiquitous, unregulated, high levels of expression, which can render rescue ineffective, and adversely affect embryonic development (325, 326). Furthermore, GFP tags in some cases alter the behaviour of recombinant proteins (327). Therefore, it is to be anticipated that injection of wildtype RNA does not always rescue a morphant phenotype. My results highlight the importance of appropriate spatial-temporal expression of *b3galnt2*.

4.4.3 The function of B3GALNT2

Very recently, a new patient with pathogenic *B3GALNT2* variants has been identified (328). This girl had dystroglycanopathy, but with a milder phenotype than the patients we described, expanding the phenotypic spectrum of this group of patients.

In addition to its role in muscle integrity, α-DG acts as a receptor by which some pathogens including Lassa virus, enter cells (329). When α-DG is hypoglycosylated, these pathogens cannot enter. A recent study elegantly exploited this fact to screen for genes that may cause dystroglycanopathy (256). The authors used mutagenised, haploid HAP1 cell lines, and identified those that were resistant to virus entry. They next identified the genes that were mutated in these cell lines, concluding that these were likely to be involved in α-DG glycosylation, and were therefore good candidates for dystroglycanopathy. Some known genes were identified, including *LARGE*, *ISPD* and *DAG1*, as were several novel genes including *B3GALNT2*. This further emphasises the importance of *B3GALNT2* in the glycosylation of α-DG, and the pathology of dystroglycanopathy.

The mannose residues initially added to the mucin-like domain of α-DG are extensively extended and branched. For example, a β1,4 linked GlcNAc groups can be added to a mannose residue by an unknown GlcNAc transferase (227). After this, a novel β1,3-N-acetylgalactosaminyltransferase acts on this structure to complete the trisaccharide GalNAcβ1-3-GlcNAc-β1,4-Man (315). This trisaccharide is required for ligand binding of α-DG. We speculated that this novel β1,3-N-acetylgalactosaminyltransferase was B3GALNT2 (314). This trisaccharide is phosphorylated and further extended by LARGE (249).

An important subsequent study used various biochemical techniques including mass spectrometry to demonstrate that GTDC2 is the unknown GlcNAc transferase mentioned above, and furthermore confirmed our speculation as to the precise function of B3GALNT2 (258). Additionally, the authors found that POMK phosphorylates the 6-position of the mannose of this trisaccharide, which is required for the activity of LARGE (Figure 4-14). Thus, the pieces of the dystroglycanopathy puzzle are beginning to fall into place.

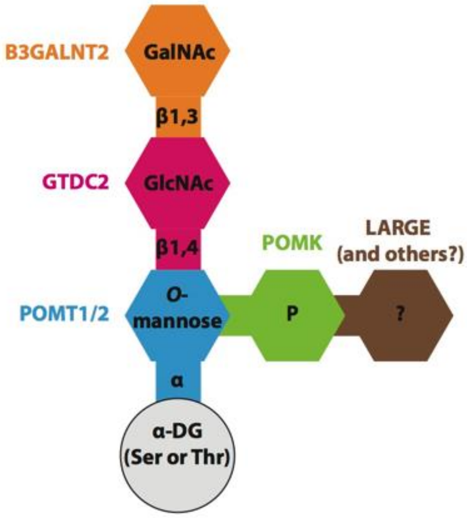

Figure 4 - 14: B3GALNT2 catalyses the synthesis of the trisaccharide GalNAcβ1-3-GlcNAc-β1,4-Man, which is required for laminin binding.
The enzyme that catalyses the addition of each monosaccharide is shown in the same colour as the monosaccharide. The precise mechanism of action of LARGE is not yet fully understood, and it may require the activity of other enzymes (258).

4.4.4 The function of GMPPB

Since the publication of our study, another group also found pathogenic variants in *GMPPB* by exome sequencing in a single family with dystroglycanopathy, brain abnormalities, and seizures, confirming the importance of *GMPPB* in the glycosylation of α-DG (330).

GDP-man is the substrate required for N-glycosylation, and also for the synthesis of Dol-P-Man, which is in turn required for all forms of glycosylation (Figure 4-2). The synthesis of Dol-P-Man is catalysed by the DPM synthase complex, which consists of DPM1, DPM2, and DPM3. One might therefore expect the phenotype of patients with pathogenic *GMPPB* variants to be more similar to those with pathogenic variants in components of the DPM synthase complex, than those with pathogenic variants in genes encoding more downstream enzymes in the pathway such as *B3GALNT2* or

POMT1. For example, patients with pathogenic variants in components of the DPM synthase complex or *DOLK* can also have defects in N-glycosylation, and an associated phenotype of CDG (233-237). However, surprisingly, patients with pathogenic *GMPPB* variants have neither defective N-glycosylation, nor signs of CDG (313).

There are several possible reasons for this. For example, Dol-P-Man might not be equally important in the N-glycosylation and O-mannosylation pathways. Previous research has shown that two enzymes in the glycosylation pathway compete for a common substrate, and can use this substrate differently, supporting this hypothesis (331). Alternatively, N-glycosylation could occur before O-mannosylation and the amount of Dol-P-Man would therefore be depleted by N-glycosylation before O-mannosylation starts. Also, the possibility of tissue-specific defects in N-glycosylation cannot be excluded.

Experiments in the pig showed that the enzyme GDP-man pyrophosphorylase consists of two subunits (317). GMPPA has GDP-Glc pyrophosphorylase activity, whereas GMPPB catalyses the formation of GDP-mannose from Mannose-1-phosphate and GTP. GMPPB has a high affinity for synthesising GDP-mannose, and a low but detectable affinity for synthesising GDP-Glc. Studies on the affinity of GMPPA to synthesise GDP-man have not been published thus far, but these results suggest some functional overlap which may be relevant to the phenotype of our patients. In humans, GMPPA and GMPPB are 30% identical.

Since the publication of our study, another group found that pathogenic variants in *GMPPA* cause a syndrome of achalasia (constriction of gastric cardia), deficiency of tear secretion, ID, gait abnormalities, neurological defects, and feeding difficulties (332). Interestingly, these patients also do not have a N-glycosylation defect. The amount of GDP-man increases in the cells of these patients, suggesting that GMPPA might negatively regulate GMPPB.

GMPPB orthologues have been knocked down in various species, including *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Arabidopsis thaliana*, *Solanum tuberosum*, *Trypanosoma brucei*, and *Leishmania mexicana* (333-339). This caused glycosylation defects and a range of pathogenic phenotypes from defective cell growth to lethality. This severity suggests that complete loss of function of *GMPPB* might be lethal. This hypothesis is supported by the fact that we did not identify any case with two null alleles (313), and also by the fact that *GMPPB* was not identified as a candidate dystroglycanopathy-associated gene by the lassa virus screen I have described (256).

4.4.5 Zebrafish phenotypes in context

With the exception of *dag1* (307), no stable germ line zebrafish mutants have been generated as models of dystroglycanopathy. Instead, most groups have used MOs. Some studies that use MO knockdown technology in zebrafish embryos to study dystroglycanopathy-associated genes focus more on some phenotypes than others. For example Buysse *et al.* do not report extensively on the gross morphology of their *b3gnt1* model, and Avsar-Ban *et al.* do not report extensively on the muscle structure of their *pomt1* and *pomt2* models (16, 306). Additionally, some studies used coinjection of *p53* MO, and others did not. Studies of *fkrp* and *gtdc2* zebrafish models reported generalised neurodegeneration, manifesting as opaque appearance of neurons in the brain, in models (17, 254). However, this observation is a well-reported non-specific effect of MOs, which is often ameliorated by the coinjection of a *p53* MO (Figure 4-6) (302, 303), which neither of these studies did. Therefore it is conceivable that this observation in these studies, which initially appears to be an interesting phenotype specific to knockdown of these genes, is in fact an artifact.

Despite these differences in methodology and reporting, some interesting insights can be gained from comparing the phenotypes of the *b3galnt2* and *gmppb* models described here to zebrafish models of six other dystroglycanopathy-associated genes (*fkrp*, *b3gnt1*, *gtdc2*, *pomt1*, *pomt2* and *ispd*) (Table 4-6). The following phenotypes are universal, in every model for which they are described: developmental delay, curved tail, impaired motility, U-shaped somites, disordered muscle fibres and hypoglycosylated α-DG. These defining characteristics of zebrafish dystroglycanopathy models are all present in my *b3galnt2* and *gmppb* models. Other phenotypes are more gene-specific. For example, a dysmorphic yolk and hypopigmentation was observed only in models of *b3galnt2*, *gmppb*, and *pomt2*, pericardial effusion is observed only in models of *b3galnt2*, *fkrp*, and *pomt2*, and micropthalmia is observed only in models of *gmppb*, *fkrp*, *gtdc2*, and *ispd*.

As I have discussed, B3GALNT2 and GTDC2 are both involved in synthesising a trisaccharide on α-DG that is essential for ligand binding (258). Given the close proximity of the activity of these proteins in the glycosylation pathway, one might assume that the phenotypes of the patients and the zebrafish models of these genes might be more similar to one another than they are to the phenotypes of patients or zebrafish models with deficiencies in genes active elsewhere in the glycosylation pathway. This does not appear to be the case (Table 4-6) (254, 314). This may be because of the methodological differences discussed, or it may be that B3GALNT2 and GTDC2 have targets in addition to α-DG, which may be different from one another. The known dystroglycanopathy-associated genes whose function is closest in the glycosylation pathway to *GMPPB* are *DPM1*, *DPM2*, *DPM3*, and *DOLK*. Zebrafish models of these genes have not yet been generated and studied. It would be interesting to see whether the phenotypes of these models would have some of the features more specific to my gmppb model, such as micropthalmia, retracted muscle fibres and damage to sarcolemma.

Gene	Reference	Developmental delay	Hydrocephalus	Microcephaly	Abnormal brain	Abnormal notochord	Impaired motility	Curved tail	Dysmorphic yolk	Hypopigmentation	Pericardial effusion	Micropthalmia	Retinal degeneration	U-shaped somites	Disordered muscle fibres	Gaps in myosepta	Lesions between muscle fibres	Retracted muscle fibres	Damage to sarcolemma	Hypoglycosylated α-DG
b3galnt2	(314)	Y	Y	N	N	X	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	N	Y
gmppb	(313)	Y	Y	N	N	X	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y
fkrp	(17, 305)	Y	N	N	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	X	Y	X	Y
b3gnt1	(16)	X	X	X	X	X	X	X	X	X	X	X	X	Y	Y	Y	Y	Y	Y	Y
gtdc2	(254)	Y	Y	Y	Y	X	Y	Y	N	N	N	Y	Y	Y	Y	X	X	X	X	Y
pomt1	(306)	Y	X	N	X	X	X	Y	N	N	N	N	N	Y	X	X	X	X	X	Y
pomt2	(306)	Y	X	N	X	X	X	Y	Y	Y	Y	N	Y	Y	X	X	X	X	X	Y
ispd	(282)	X	Y	N	Y	X	Y	X	X	X	X	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 4 - 6: Phenotypic comparison of zebrafish dystroglycanopathy models.
All models made by MO knockdown. Y = phenotype observed; N = phenotype not observed; X = phenotype not investigated or described.

4.4.6 Technical limitations of this study

There are technical challenges associated with the use of MOs, as discussed in section 4.1.5. In this study, I took various measures to limit the effect of these challenges on my results. For example, I coinjected a *p53* MO along with the MOs for my genes of interest, in order to limit any generalised p53 upregulation, which may have produced non-specific results (302, 303). I also confirmed that the *b3galnt2* and *gmppb* MOs inhibit the expression of *b3galnt2* and *gmppb*. Despite the technical limitations, MOs have hitherto been the method of choice for generating zebrafish models of dystroglycanopathy. The overlap between the phenotypes I observed in my two models, and the phenotypes observed in other dystroglycanopathy models, support the hypothesis that those phenotypes are specific to dystroglycanopathy models, as discussed in section 4.4.5.

Nevertheless, it is likely that in coming years the method of choice for generating zebrafish models of dystroglycanopathy will move towards CRISPRs. CRISPRs are cheap, easy to generate, and allow very specific, targeted mutations to be introduced into the genome (300). While the experiments would take longer than for MOs, this disadvantage would be outweighed by the fact that the results are likely to be much more consistent and specific. Technological advances continue to improve the specificity of CRISPRs (340). The existence of endogenous CRISPRs as an immune-like mechanism is prokaryotes has been known for years (341), however their use as a genome-editing tool was not developed until 2012, and are still in the process of being optimised and becoming widely available (342). Had CRISPRs been available at the time this study was designed (2011), it is probable that we would have elected to use them rather than MOs to generate zebrafish models of *B3GALNT2* and *GMPPB*.

4.4.7 Future research

A further experiment that could be performed is to attempt to rescue the phenotype of the *GMPPB* morphants by coinjection of wildtype human or zebrafish *GMPPB* RNA. Unfortunately, it was not possible to carry out this experiment in the time frame of this project. While a positive result from this experiment would have provided useful supportive evidence that the phenotypes we observed in the *GMPPB* morphants were specific to knockdown of *GMPPB*, my colleagues and I did not feel that this experiment was essential for the following reasons. First, as discussed in section 4.4.2, coinjection of wildtype RNA does not always result in phenotypic rescue. Second, even without this experiment, we were confident in our conclusion that *GMPPB* variants in the patients were causative of disease, because the phenotype of the *gmppb* morphants was very consistent with those of the patients and those of other zebrafish models of dystroglycanopathy-associated genes, and because of all the other clinical and cellular evidence.

In the wider field of dystroglycanopathy research, many questions remain. Indeed, identification of the genes associated with disease is only the first step. Elucidating the precise role of each gene in the glycosylation of α-DG is vital to understand disease pathology at the molecular level. Here, the collaboration of clinicians, human geneticists, and model organism researchers with biochemists will prove fruitful. For example, Yoshida-Moriguchi *et al.* have shed light on the exact function of several dystroglycanopathy-associated genes, including *B3GALNT2* (258).

Several observations suggest that some dystroglycanopathy-associated genes may be involved in the glycosylation of other target proteins in addition to α-DG. These include the heterogeneity of dystroglycanopathy (both in terms of phenotypic severity and organ systems affected), the lack of correlation between the extent to which α-DG is hypoglycosylated, and the clinical severity of disease (292), and the known promiscuity of some bacterial glycosyltransferases (343-345). Identifying any other target proteins of dystroglycanopathy-associated genes would undoubtedly improve understanding of the disease. Similarly, identification of secondary modifiers may help to explain why some patients are less severely affected than others, and may point towards therapeutic targets.

Therapeutic sugar supplementation can treat patients with some glycosylation defects. The best example of this is CDG caused by pathogenic variants in the gene encoding mannose phosphate isomerase, which catalyses the conversion of fructose-6-phosphate to mannose-6-phosphate. The resulting deficiency in mannose-6-phosphate (which is a precursor to mannose-1-phosphate, the substrate of GMPPB) causes a multisystem phenotype including gastrointestinal and liver disease (346). Orally administrated mannose can improve clinical symptoms of these patients, and the levels of glycosylation of glycoproteins (347). This works because mannose can be converted to mannose-6-phosphate, catalysed by hexokinase. This is normally a minor alternative pathway, but is promoted in the presence of high doses of exogenous mannose.

Mannose supplementation in water improves the phenotype of a zebrafish MO model of CDG-1b (348). This suggests that a similar approach, using supplementation of mannose or GDP-mannose, could be a therapeutic avenue worth exploring in patients with pathogenic *GMPPB* variants. Furthermore, a zebrafish model could be a useful initial tool to test the viability of this idea. It is clear that the zebrafish will continue going from strength to strength as a model of dystroglycanopathy.

5 Discussion

In this dissertation I have described three projects that use next generation sequencing (NGS) to identify variants that can cause rare developmental disorders, along with statistical or functional follow-up approaches. The aim of the project described in chapter 2 was to explore how well exome sequencing performs as a method for identifying variants that cause abnormal fetal development. Exome sequencing of 30 parent-fetus trios was performed, where the fetuses had structural abnormalities. I identified single nucleotide variants (SNVs), insertion deletions (indels), and copy number variants (CNVs) with *de novo*, autosomal recessive, or X-linked (for male fetuses) inheritance in this cohort. I investigated various methods of variant prioritisation and interpretation, and concluded that for 3/30 fetuses (10%) a causal mutation had been identified. All of these were *de novo*, emphasising the importance of sequencing trios, and showing that there is a low recurrence risk for future pregnancies of these couples. Only one of these three mutations was a CNV and could therefore have been detected by microarray, the highest resolution genome-wide method currently used in prenatal genetic diagnostics. No novel disease-associated genes were identified during this study, because it was underpowered for this due to the small cohort size, and diversity of the fetal phenotypes. Nevertheless, this study demonstrates the utility of exome sequencing for prenatal genetic diagnosis, and paves the way for similar, larger studies. Issues that would need to be addressed before exome sequencing could become widely used for prenatal genetic diagnostics include the development and implementation of a primarily computational variant interpretation pipeline, and resolution of some contentious ethical issues. Based on this project, it seems clear that NGS is the future of prenatal diagnostics.

In chapter 3, I described a targeted resequencing study that was performed on a cohort of patients with intellectual disability (ID) as part of the UK10K project. I designed and implemented an analytical pipeline to identify variants that were likely to be causative. The first aim of this project was to identify causal variants in known ID-associated genes in the cohort. Using my pipeline, and further interpretation of variants by clinical collaborators, likely causative variants in known ID-associated genes were found for 14% of the cohort. The second aim was to attempt to identify any novel ID-associated genes. We found causative *de novo* loss of function mutations in the putative histone methyltransferase gene *SETD5* in seven patients with ID, and showed that loss of function of *SETD5* is probably responsible for many features of 3p25 microdeletion syndrome, as well as being a relatively common cause of sporadic ID. This finding also emphasises the importance of methyltransferases in the pathology of ID. The final aim of the targeted resequencing study described in chapter 3 was to ascertain whether there is a burden of variants in ID-associated genes in ID patients compared to controls. I used the cohort allelic sums test to demonstrate that there is a burden of both loss of function variants, and some categories of missense variants, in ID-associated genes in ID patients compared to controls. This project demonstrates the importance of rigorous statistical methods in assigning causality to a gene associated with a rare developmental disorder. It also shows how case-control enrichment analyses can be a valuable statistical follow-up approaches to NGS, as it can focus attention on specific classes of variant with a higher likelihood of being pathogenic.

The aims of the project described in chapter 4 were to make zebrafish models of dystroglycanopathy using morpholino oligonucleotides to inhibit the expression of the two candidate dystroglycanopathy-associated genes *B3GALNT2* and *GMPPB*, and to determine the extent to which the phenotype of these models recapitulated the phenotypes of the patients. I first showed that zebrafish embryos are appropriate models for *B3GALNT2* and *GMPPB*, and that morpholinos do inhibit their expression. I next used several assays including immunofluorescence staining, Evans blue dye, and immunoblotting to determine the phenotype of the models compared to wildtype embryos. I found that there were similarities between the zebrafish models and the patients in terms of gross appearance and behaviour (such as movement defects), muscle structure (such as disordered fibres), and molecular level (hypoglycosylation of α-DG). This phenotype data from these two zebrafish models, together with clinical patient data and cellular models, led to the conclusion that variants in *B3GALNT2* and *GMPPB* can indeed cause dystroglycanopathy.

The work described in this dissertation has four important outcomes that I think directly or indirectly could improve the lives of patients affected by rare developmental disorders. All of this work was done in association with colleagues, collaborators and supervisors, as described in the Acknowledgements section and at the relevant portions of the text. First, a genetic cause was identified for 10% of the cohort of 30 fetuses with structural abnormalities, and for 14% of the UK10K ID cohort. Where the results were returned to the families, this ended the ‘diagnostic odyssey’ for them, and could allow their clinicians to estimate recurrence risk for future pregnancies. Furthermore, it revealed some insights into the pathology of rare developmental disorders, for example the importance of *de novo* mutations in abnormal fetal development. Second, we demonstrated that exome sequencing is a promising tool for prenatal genetic diagnostics, and may be better than the current gold-standard method. This cohort of 30 fetuses with structural abnormalities is the largest such cohort published to date, nonetheless, my findings represent a 'proof-of-principle' study that paves the way for the even larger-scale evaluations of NGS for prenatal genetic diagnosis that are needed to both accurately quantify the diagnostic yield and identify novel genetic causes of fetal abnormalities. Third, the case-control enrichment analyses of ID patients in the UK10K study revealed some interesting findings into the genetic architecture of ID, many of which support findings that have previously been shown by other methods. For example, *de novo* mutations are an important cause of ID, and there is a burden of variants in some candidate genes that have not yet been conclusively associated with ID. The most interesting and novel finding from these analyses was that there is an enrichment of certain categories of missense variant, such as predicted-damaging X-linked variants, in ID-associated genes in ID patients compared to controls. Finally, and perhaps most importantly, the work described in this dissertation has contributed towards the discovery of three novel developmental disease-associated genes: *B3GALNT2* and *GMPPB* in dystroglycanopathy, and *SETD5* in ID. This will have a direct impact upon patients who have disease caused by damaging variants in these genes, as now they are more likely to receive a genetic diagnosis. Additionally, it improves understanding of the disease, for example *SETD5* emphasises the importance of appropriate histone methylation in normal cognitive functioning.

In conclusion, the three projects described in this dissertation highlight the importance of NGS for understanding rare developmental disorders. NGS, whether it is exome sequencing, whole genome sequencing, or targeted resequencing of candidate genes, has proved to be a valuable tool for clinical diagnosis of rare developmental disorders, and for the discovery of novel disease-associated genes. Often, statistical or functional follow-up approaches are required to confirm that variants in a particular gene do cause the disorder. An increasing number of genes that are associated with rare developmental disorders are being identified through the use of NGS. As progress continues to be made in this area, the focus of the research community is likely to shift towards understanding the precise mechanisms by which variants in a given gene cause a rare developmental disorder, so that ultimately therapies for these disorders might be developed. It is likely that statistical follow-up approaches such as case-control enrichment analyses, and functional follow-up approaches such as modelling candidate genes in an organism such as the zebrafish, will be valuable for increasing this understanding. It is clear that NGS, along with supplementary and follow-up approaches, are both directly and indirectly improving the lives of patients with rare developmental disorders, and will continue to do so for the foreseeable future.

6 References

1. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature reviews Genetics. 2013 Oct;14(10):681-91. PubMed PMID: 23999272.

2. Rilstone JJ, Alkhater RA, Minassian BA. Brain dopamine-serotonin vesicular transport disease and its treatment. N Engl J Med. 2013 Feb 7;368(6):543-50. PubMed PMID: 23363473.

3. Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP, et al. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. Am J Hum Genet. 2014 Jun 5;94(6):809-17. PubMed PMID: 24906018.

4. Wu CH, Fallini C, Ticozzi N, Keagle PJ, Sapp PC, Piotrowska K, et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. Nature. 2012 Aug 23;488(7412):499-503. PubMed PMID: 22801503. Pubmed Central PMCID: 3575525.

5. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat Genet. 2012 Apr;44(4):376-8. PubMed PMID: 22426308.

6. Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, et al. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science. 1985 Nov 29;230(4729):1054-7. PubMed PMID: 2997931.

7. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010 Jan;42(1):30-5. PubMed PMID: 19915526. Pubmed Central PMCID: 2847889.

8. Wang Z, Liu X, Yang BZ, Gelernter J. The role and challenges of exome sequencing in studies of human diseases. Frontiers in genetics. 2013;4:160. PubMed PMID: 24032039. Pubmed Central PMCID: 3752524.

9. Firth HV, Wright CF. The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol. 2011 Aug;53(8):702-3. PubMed PMID: 21679367. Epub 2011/06/18. eng.

10. Kaye J, Hurles M, Griffin H, Grewal J, Bobrow M, Timpson N, et al. Managing clinically significant findings in research: the UK10K example. Eur J Hum Genet. 2014 Jan 15. PubMed PMID: 24424120. Pubmed Central PMCID: 4026295.

11. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013 Oct 17;369(16):1502-11. PubMed PMID: 24088041.

12. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014 Apr 24;508(7497):469-76. PubMed PMID: 24759409.

13. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet. 2009 May;41(5):535-43. PubMed PMID: 19377476. Pubmed Central PMCID: 2872007. Epub 2009/04/21. eng.

14. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014 Feb 13;506(7487):185-90. PubMed PMID: 24463508.

15. Bhattacharya S, Das A, Ghosh S, Dasgupta R, Bagchi A. Hypoglycosylation of dystroglycan due to T192M mutation: a molecular insight behind the fact. Gene. 2014 Mar 1;537(1):108-14. PubMed PMID: 24361964.

16. Buysse K, Riemersma M, Powell G, van Reeuwijk J, Chitayat D, Roscioli T, et al. Missense mutations in beta-1,3-N-acetylglucosaminyltransferase 1 (B3GNT1) cause Walker-Warburg syndrome. Hum Mol Genet. 2013 Jan 28. PubMed PMID: 23359570. Epub 2013/01/30. Eng.

17. Thornhill P, Bassett D, Lochmuller H, Bushby K, Straub V. Developmental defects in a zebrafish model for muscular dystrophies associated with the loss of fukutin-related protein (FKRP). Brain : a journal of neurology. 2008 Jun;131(Pt 6):1551-61. PubMed PMID: 18477595. Epub 2008/05/15. eng.

18. Springett A, Morris JK. Congenital Anomaly Statistics. England and Wales. London: British Isles Network of Congenital Anomaly Registers. 2010.

19. Vlastos IM, Koudoumnakis E, Houlakis M, Nasika M, Griva M, Stylogianni E. Cleft lip and palate treatment of 530 children over a decade in a single centre. Int J Pediatr Otorhinolaryngol. 2009 Jul;73(7):993-7. PubMed PMID: 19443049.

20. Verity C, Firth H, ffrench-Constant C. Congenital abnormalities of the central nervous system. J Neurol Neurosurg Psychiatry. 2003 Mar;74 Suppl 1:i3-8. PubMed PMID: 12611928. Pubmed Central PMCID: 1765611.

21. Cereda A, Carey JC. The trisomy 18 syndrome. Orphanet J Rare Dis. 2012;7:81. PubMed PMID: 23088440. Pubmed Central PMCID: 3520824.

22. Hillman SC, McMullan DJ, Hall G, Togneri FS, James N, Maher EJ, et al. Use of prenatal chromosomal microarray: prospective cohort study and systematic review and meta-analysis. Ultrasound Obstet Gynecol. 2013 Jun;41(6):610-20. PubMed PMID: 23512800. Epub 2013/03/21. eng.

23. Rousseau F, el Ghouzzi V, Delezoide AL, Legeai-Mallet L, Le Merrer M, Munnich A, et al. Missense FGFR3 mutations create cysteine residues in thanatophoric dwarfism type I (TD1). Hum Mol Genet. 1996 Apr;5(4):509-12. PubMed PMID: 8845844. Epub 1996/04/01. eng.

24. Tabor A, Vestergaard CH, Lidegaard O. Fetal loss rate after chorionic villus sampling and amniocentesis: an 11-year national registry study. Ultrasound Obstet Gynecol. 2009 Jul;34(1):19-24. PubMed PMID: 19504504.

25. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. Lancet. 1997 Aug 16;350(9076):485-7. PubMed PMID: 9274585.

26. Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, et al. Chemical differentiation along metaphase chromosomes. Exp Cell Res. 1968 Jan;49(1):219-22. PubMed PMID: 5640698.

27. Shaffer LG, Bejjani BA. A cytogeneticist's perspective on genomic microarrays. Hum Reprod Update. 2004 May-Jun;10(3):221-6. PubMed PMID: 15140869.

28. Bauman JG, Wiegant J, Borst P, van Duijn P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. Exp Cell Res. 1980 Aug;128(2):485-90. PubMed PMID: 6157553.

29. Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, et al. Multicolor spectral karyotyping of human chromosomes. Science. 1996 Jul 26;273(5274):494-7. PubMed PMID: 8662537.

30. Evans MI, Henry GP, Miller WA, Bui TH, Snidjers RJ, Wapner RJ, et al. International, collaborative assessment of 146,000 prenatal karyotypes: expected limitations if only chromosome-specific probes and fluorescent in-situ hybridization are used. Hum Reprod. 1999 May;14(5):1213-6. PubMed PMID: 10325263.

31. Nickerson E, Greenberg F, Keating MT, McCaskill C, Shaffer LG. Deletions of the elastin gene at 7q11.23 occur in approximately 90% of patients with Williams syndrome. Am J Hum Genet. 1995 May;56(5):1156-61. PubMed PMID: 7726172. Pubmed Central PMCID: 1801441.

32. Ligon AH, Kashork CD, Richards CS, Shaffer LG. Identification of female carriers for Duchenne and Becker muscular dystrophies using a FISH-based approach. Eur J Hum Genet. 2000 Apr;8(4):293-8. PubMed PMID: 10854113.

33. Brackley KJ, Kilby MD, Morton J, Whittle MJ, Knight SJ, Flint J. A case of recurrent congenital fetal anomalies associated with a familial subtelomeric translocation. Prenat Diagn. 1999 Jun;19(6):570-4. PubMed PMID: 10416976.

34. Wapner RJ, Martin CL, Levy B, Ballif BC, Eng CM, Zachary JM, et al. Chromosomal microarray versus karyotyping for prenatal diagnosis. N Engl J Med. 2012 Dec 6;367(23):2175-84. PubMed PMID: 23215555. Epub 2012/12/12. eng.

35. Hillman SC, Pretlove S, Coomarasamy A, McMullan DJ, Davison EV, Maher ER, et al. Additional information from array comparative genomic hybridization technology over conventional karyotyping in prenatal diagnosis: a systematic review and meta-analysis. Ultrasound Obstet Gynecol. 2011 Jan;37(1):6-14. PubMed PMID: 20658510.

36. Srebniak MI, Boter M, Oudesluijs GO, Cohen-Overbeek T, Govaerts LC, Diderich KE, et al. Genomic SNP array as a gold standard for prenatal diagnosis of foetal ultrasound abnormalities. Molecular cytogenetics. 2012;5(1):14. PubMed PMID: 22413963. Pubmed Central PMCID: 3328283.

37. Vanakker O, Vilain C, Janssens K, Van der Aa N, Smits G, Bandelier C, et al. Implementation of genomic arrays in prenatal diagnosis: The Belgian approach to meet the challenges. Eur J Med Genet. 2014 Feb 15. PubMed PMID: 24534801.

38. Hillman SC, McMullan DJ, Maher ER, Kilby MD. The use of chromosomal microarray in prenatal diagnosis. The Obstetrician and Gynaecologist. 2013;15(2):80-4.

39. UKGTN. UK Genetic Testing Network arrayCGH Commissioning Workshop. London: phg foundation: 2009.

40. Valduga M, Philippe C, Bach Segura P, Thiebaugeorges O, Miton A, Beri M, et al. A retrospective study by oligonucleotide array-CGH analysis in 50 fetuses with multiple malformations. Prenat Diagn. 2010 Apr;30(4):333-41. PubMed PMID: 20155755. Epub 2010/02/16. eng.

41. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011 Jul;43(7):712-4. PubMed PMID: 21666693. Pubmed Central PMCID: 3322360. Epub 2011/06/15. eng.

42. Mann K, Ogilvie CM. QF-PCR: application, overview and review of the literature. Prenat Diagn. 2012 Apr;32(4):309-14. PubMed PMID: 22467160.

43. Lo YM, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM, et al. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am J Hum Genet. 1998 Apr;62(4):768-75. PubMed PMID: 9529358. Pubmed Central PMCID: 1377040.

44. Lun FM, Chiu RW, Allen Chan KC, Yeung Leung T, Kin Lau T, Dennis Lo YM. Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clin Chem. 2008 Oct;54(10):1664-72. PubMed PMID: 18703764.

45. Palomaki GE, Deciu C, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, et al. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med. 2012 Mar;14(3):296-305. PubMed PMID: 22281937. Pubmed Central PMCID: 3938175.

46. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. Clin Chem. 2010 Aug;56(8):1279-86. PubMed PMID: 20558635.

47. Nicolaides KH, Syngelaki A, Gil M, Atanasova V, Markova D. Validation of targeted sequencing of single-nucleotide polymorphisms for non-invasive prenatal

detection of aneuploidy of chromosomes 13, 18, 21, X, and Y. Prenat Diagn. 2013 Jun;33(6):575-9. PubMed PMID: 23613152.

48. Ge H, Huang X, Li X, Chen S, Zheng J, Jiang H, et al. Noninvasive prenatal detection for pathogenic CNVs: the application in alpha-thalassemia. PLoS One. 2013;8(6):e67464. PubMed PMID: 23840709. Pubmed Central PMCID: 3696090.

49. Lim JH, Kim MJ, Kim SY, Kim HO, Song MJ, Kim MH, et al. Non-invasive prenatal detection of achondroplasia using circulating fetal DNA in maternal plasma. J Assist Reprod Genet. 2011 Feb;28(2):167-72. PubMed PMID: 20963478. Pubmed Central PMCID: 3059531.

50. Agarwal A, Sayres LC, Cho MK, Cook-Deegan R, Chandrasekharan S. Commercial landscape of noninvasive prenatal testing in the United States. Prenat Diagn. 2013 Jun;33(6):521-31. PubMed PMID: 23686656. Pubmed Central PMCID: 3898859.

51. Song Y, Liu C, Qi H, Zhang Y, Bian X, Liu J. Noninvasive prenatal testing of fetal aneuploidies by massively parallel sequencing in a prospective Chinese population. Prenat Diagn. 2013 Jul;33(7):700-6. PubMed PMID: 23703459.

52. Hill M, Karunaratna M, Lewis C, Forya F, Chitty L. Views and preferences for the implementation of non-invasive prenatal diagnosis for single gene disorders from health professionals in the United Kingdom. Am J Med Genet A. 2013 Jul;161A(7):1612-8. PubMed PMID: 23696422.

53. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. Non-invasive prenatal measurement of the fetal genome. Nature. 2012 Jul 19;487(7407):320-4. PubMed PMID: 22763444. Pubmed Central PMCID: 3561905.

54. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, et al. Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med. 2012 Jun 6;4(137):137ra76. PubMed PMID: 22674554. Pubmed Central PMCID: 3379884.

55. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nature reviews Genetics. 2012;13(8):565-75. PubMed PMID: 22805709. Epub 2012/07/19. eng.

56. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011 Jun;43(6):585-9. PubMed PMID: 21572417. eng.

57. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012 Nov 15;367(20):1921-9. PubMed PMID: 23033978. Epub 2012/10/05. eng.

58. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013 Jun 13;498(7453):220-3. PubMed PMID: 23665959. Pubmed Central PMCID: 3706629. Eng.

59. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10;461(7261):272-6. PubMed PMID: 19684571. Pubmed Central PMCID: 2844771.

60. Glockle N, Kohl S, Mohr J, Scheurenbrand T, Sprecher A, Weisschuh N, et al. Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies. Eur J Hum Genet. 2014 Jan;22(1):99-104. PubMed PMID: 23591405. Pubmed Central PMCID: 3865404.

61. Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet Med. 2012 Jan;14(1):51-9. PubMed PMID: 22237431.

62. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014 Jun 4. PubMed PMID: 24896178.

63. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 2014 Jul 19;15(1):247. PubMed PMID: 25038816.

64. Piton A, Redin C, Mandel JL. XLID-Causing Mutations and Associated Genes Challenged in Light of Data From Large-Scale Human Exome Sequencing. Am J Hum Genet. 2013 Jul 18. PubMed PMID: 23871722. Pubmed Central PMCID: 3738825.

65. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9. PubMed PMID: 20354512. Pubmed Central PMCID: 2855889.

66. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81. PubMed PMID: 19561590.

67. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 Jul;15(7):901-13. PubMed PMID: 15965027. Pubmed Central PMCID: 1172034.

68. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS genetics. 2010 Oct;6(10):e1001154. PubMed PMID: 20976243. Pubmed Central PMCID: 2954820.

69. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006 May;24(5):537-44. PubMed PMID: 16680138. Epub 2006/05/09. eng.

70. Dan S, Chen F, Choy KW, Jiang F, Lin J, Xuan Z, et al. Prenatal detection of aneuploidy and imbalanced chromosomal arrangements by massively parallel sequencing. PLoS One. 2012;7(2):e27835. PubMed PMID: 22389664. Pubmed Central PMCID: 3289612.

71. Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, et al. Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med. 2012 Dec 6;367(23):2226-32. PubMed PMID: 23215558. Epub 2012/12/12. eng.

72. Filges I, Nosova E, Bruder E, Tercanli S, Townsend K, Gibson W, et al. Exome sequencing identifies mutations in KIF14 as a novel cause of an autosomal recessive lethal fetal ciliopathy phenotype. Clin Genet. 2013 Oct 15. PubMed PMID: 24128419.

73. Carss KJ, Hillman SC, Parthiban V, McMullan DJ, Maher ER, Kilby MD, et al. Exome sequencing improves genetic diagnosis of structural fetal abnormalities revealed by ultrasound. Hum Mol Genet. 2014 Feb 11. PubMed PMID: 24476948.

74. Mackie FL, Carss KJ, Hillman SC, Hurles ME, Kilby MD. Exome sequencing in fetuses with structural malformations [review]. Journal of Clinical Medicine. 2014;3(3):747-62.

75. Robinson PN, Mundlos S. The human phenotype ontology. Clin Genet. 2010 Jun;77(6):525-34. PubMed PMID: 20412080.

76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. PubMed PMID: 19505943. Pubmed Central PMCID: 2723002.

77. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011 Jun;21(6):961-73. PubMed PMID: 20980555. Pubmed Central PMCID: 3106329.

78. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8. PubMed PMID: 21653522. Pubmed Central PMCID: 3137218.

79. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D48-55. PubMed PMID: 23203987. Pubmed Central PMCID: 3531136.

80. Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. PubMed PMID: 23128226. Pubmed Central PMCID: 3498066.

81. Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013 Oct;10(10):985-7. PubMed PMID: 23975140.

82. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178-92. PubMed PMID: 22517427. Pubmed Central PMCID: 3603213.

83. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007 Apr;80(4):727-39. PubMed PMID: 17357078. Pubmed Central PMCID: 1852724. Epub 2007/03/16. eng.

84. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012 Sep 26. PubMed PMID: 23020937. Epub 2012/10/02. Eng.

85. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000 Jan 1;28(1):352-5. PubMed PMID: 10592272. Pubmed Central PMCID: 102496.

86. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet. 2013 Oct 3;93(4):631-40. PubMed PMID: 24055113. Pubmed Central PMCID: 3791261.

87. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012 Jan;71(1):5-14. PubMed PMID: 22275248.

88. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011;12(7):R68. PubMed PMID: 21787409. Pubmed Central PMCID: 3218830.

89. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006 Sep;16(9):1182-90. PubMed PMID: 16902084. Pubmed Central PMCID: 1557762.

90. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, et al. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am J Hum Genet. 2010 Sep 10;87(3):316-24. PubMed PMID: 20797689. Pubmed Central PMCID: 2933353. Epub 2010/08/28. eng.

91. Foldynova-Trantirkova S, Wilcox WR, Krejci P. Sixteen years and counting: the current understanding of fibroblast growth factor receptor 3 (FGFR3) signaling in skeletal dysplasias. Hum Mutat. 2012 Jan;33(1):29-41. PubMed PMID: 22045636. Pubmed Central PMCID: 3240715. Epub 2011/11/03. eng.

92. Potocki L, Abuelo DN, Oyer CE. Cardiac malformation in two infants with hypochondrogenesis. Am J Med Genet. 1995 Nov 20;59(3):295-9. PubMed PMID: 8599352. Epub 1995/11/20. eng.

93. Rittler M, Orioli IM. Achondrogenesis type II with polydactyly. Am J Med Genet. 1995 Nov 6;59(2):157-60. PubMed PMID: 8588578. Epub 1995/11/06. eng.

94. Zankl A, Zabel B, Hilbert K, Wildhardt G, Cuenot S, Xavier B, et al. Spondyloperipheral dysplasia is caused by truncating mutations in the C-propeptide of COL2A1. Am J Med Genet A. 2004 Aug 30;129A(2):144-8. PubMed PMID: 15316962. Epub 2004/08/19. eng.

95. Nishimura G, Haga N, Kitoh H, Tanaka Y, Sonoda T, Kitamura M, et al. The phenotypic spectrum of COL2A1 mutations. Hum Mutat. 2005 Jul;26(1):36-43. PubMed PMID: 15895462. Epub 2005/05/17. eng.

96. De Luca A, Bottillo I, Sarkozy A, Carta C, Neri C, Bellacchio E, et al. NF1 gene mutations represent the major molecular event underlying neurofibromatosis-Noonan syndrome. Am J Hum Genet. 2005 Dec;77(6):1092-101. PubMed PMID: 16380919. Pubmed Central PMCID: 1285166. Epub 2005/12/29. eng.

97. Fahsold R, Hoffmeyer S, Mischung C, Gille C, Ehlers C, Kucukceylan N, et al. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. Am J Hum Genet. 2000 Mar;66(3):790-818. PubMed PMID: 10712197. Pubmed Central PMCID: 1288164. Epub 2000/03/11. eng.

98. Padmanabhan A, Lee JS, Ismat FA, Lu MM, Lawson ND, Kanki JP, et al. Cardiac and vascular functions of the zebrafish orthologues of the type I neurofibromatosis gene NFI. Proc Natl Acad Sci U S A. 2009 Dec 29;106(52):22305-10. PubMed PMID: 19966217. Pubmed Central PMCID: 2799742. Epub 2009/12/08. eng.

99. Yan Z, Wang Z, Sharova L, Sharov AA, Ling C, Piao Y, et al. BAF250B-associated SWI/SNF chromatin-remodeling complex is required to maintain undifferentiated mouse embryonic stem cells. Stem Cells. 2008 May;26(5):1155-65. PubMed PMID: 18323406. Pubmed Central PMCID: 2409195. Epub 2008/03/08. eng.

100. Boerkoel CF, Takashima H, John J, Yan J, Stankiewicz P, Rosenbarker L, et al. Mutant chromatin remodeling protein SMARCAL1 causes Schimke immuno-osseous dysplasia. Nat Genet. 2002 Feb;30(2):215-20. PubMed PMID: 11799392. Epub 2002/01/19. eng.

101. Endele S, Rosenberger G, Geider K, Popp B, Tamer C, Stefanova I, et al. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. Nat Genet. 2010 Nov;42(11):1021-6. PubMed PMID: 20890276. Epub 2010/10/05. eng.

102. Gao L, Macara IG, Joberty G. Multiple splice variants of Par3 and of a novel related gene, Par3L, produce proteins with different binding properties. Gene. 2002 Jul 10;294(1-2):99-107. PubMed PMID: 12234671.

103. Hong E, Jayachandran P, Brewster R. The polarity protein Pard3 is required for centrosome positioning during neurulation. Dev Biol. 2010 May 15;341(2):335-45. PubMed PMID: 20138861. Pubmed Central PMCID: 2862117.

104. Hirose T, Karasawa M, Sugitani Y, Fujisawa M, Akimoto K, Ohno S, et al. PAR3 is essential for cyst-mediated epicardial development by establishing apical cortical domains. Development. 2006 Apr;133(7):1389-98. PubMed PMID: 16510507.

105. Bisschoff IJ, Zeschnigk C, Horn D, Wellek B, Riess A, Wessels M, et al. Novel mutations including deletions of the entire OFD1 gene in 30 families with type 1 orofaciodigital syndrome: a study of the extensive clinical variability. Hum Mutat. 2013 Jan;34(1):237-47. PubMed PMID: 23033313. Epub 2012/10/04. eng.

106. Durkin ME, Avner MR, Huh CG, Yuan BZ, Thorgeirsson SS, Popescu NC. DLC-1, a Rho GTPase-activating protein with tumor suppressor function, is essential for embryonic development. FEBS Lett. 2005 Feb 14;579(5):1191-6. PubMed PMID: 15710412. Epub 2005/02/16. eng.

107. Vilhais-Neto GC, Maruhashi M, Smith KT, Vasseur-Cognet M, Peterson AS, Workman JL, et al. Rere controls retinoic acid signalling and somite bilateral symmetry. Nature. 2010 Feb 18;463(7283):953-7. PubMed PMID: 20164929. Epub 2010/02/19. eng.

108. Plaster N, Sonntag C, Schilling TF, Hammerschmidt M. REREa/Atrophin-2 interacts with histone deacetylase and Fgf8 signaling to regulate multiple processes of zebrafish development. Dev Dyn. 2007 Jul;236(7):1891-904. PubMed PMID: 17576618. Epub 2007/06/20. eng.

109. Kim BJ, Zaveri HP, Shchelochkov OA, Yu Z, Hernandez-Garcia A, Seymour ML, et al. An allelic series of mice reveals a role for RERE in the development of multiple organs affected in chromosome 1p36 deletions. PLoS One. 2013;8(2):e57460. PubMed PMID: 23451234. Pubmed Central PMCID: 3581587. Epub 2013/03/02. eng.

110. Liu W, Morito D, Takashima S, Mineharu Y, Kobayashi H, Hitomi T, et al. Identification of RNF213 as a susceptibility gene for moyamoya disease and its possible role in vascular development. PLoS One. 2011;6(7):e22542. PubMed PMID: 21799892. Pubmed Central PMCID: 3140517. Epub 2011/07/30. eng.

111. Kida Y, Maeda Y, Shiraishi T, Suzuki T, Ogura T. Chick Dach1 interacts with the Smad complex and Sin3a to control AER formation and limb development along the proximodistal axis. Development. 2004 Sep;131(17):4179-87. PubMed PMID: 15280207. Epub 2004/07/29. eng.

112. Martini SR, Davis RL. The dachshund gene is required for the proper guidance and branching of mushroom body axons in Drosophila melanogaster. J Neurobiol. 2005 Aug;64(2):133-44. PubMed PMID: 15818552. Epub 2005/04/09. eng.

113. Backman M, Machon O, Van Den Bout CJ, Krauss S. Targeted disruption of mouse Dach1 results in postnatal lethality. Dev Dyn. 2003 Jan;226(1):139-44. PubMed PMID: 12508235. Epub 2003/01/01. eng.

114. Boyd PA, Keeling JW, Lindenbaum RH. Fraser syndrome (cryptophthalmos-syndactyly syndrome): a review of eleven cases with postmortem findings. Am J Med Genet. 1988 Sep;31(1):159-68. PubMed PMID: 2851937. Epub 1988/09/01. eng.

115. Pitera JE, Scambler PJ, Woolf AS. Fras1, a basement membrane-associated protein mutated in Fraser syndrome, mediates both the initiation of the mammalian kidney and the integrity of renal glomeruli. Hum Mol Genet. 2008 Dec 15;17(24):3953-64. PubMed PMID: 18787044. Pubmed Central PMCID: 2638576. Epub 2008/09/13. eng.

116. Kerecuk L, Long DA, Ali Z, Anders C, Kolatsi-Joannou M, Scambler PJ, et al. Expression of Fraser syndrome genes in normal and polycystic murine kidneys. Pediatr Nephrol. 2012 Feb 1. PubMed PMID: 22294133. Epub 2012/02/02. Eng.

117. Vrontou S, Petrou P, Meyer BI, Galanopoulos VK, Imai K, Yanagi M, et al. Fras1 deficiency results in cryptophthalmos, renal agenesis and blebbed phenotype in mice. Nat Genet. 2003 Jun;34(2):209-14. PubMed PMID: 12766770.

118. Talbot JC, Walker MB, Carney TJ, Huycke TR, Yan YL, BreMiller RA, et al. fras1 shapes endodermal pouch 1 and stabilizes zebrafish pharyngeal skeletal development. Development. 2012 Aug;139(15):2804-13. PubMed PMID: 22782724. Pubmed Central PMCID: 3392706. Epub 2012/07/12. eng.

119. Gautier P, Naranjo-Golborne C, Taylor MS, Jackson IJ, Smyth I. Expression of the fras1/frem gene family during zebrafish development and fin morphogenesis. Dev Dyn. 2008 Nov;237(11):3295-304. PubMed PMID: 18816440. Epub 2008/09/26. eng.

120. Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nat Methods. 2013 Nov;10(11):1083-4. PubMed PMID: 24076761.

121. Smedley D, Oellrich A, Kohler S, Ruef B, Sanger Mouse Genetics P, Westerfield M, et al. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database (Oxford). 2013;2013:bat025. PubMed PMID: 23660285. Pubmed Central PMCID: 3649640.

122. Woodbine L, Neal JA, Sasi NK, Shimada M, Deem K, Coleman H, et al. PRKDC mutations in a SCID patient with profound neurological abnormalities. J Clin Invest. 2013 Jul 1;123(7):2969-80. PubMed PMID: 23722905.

123. Goryunov D, He CZ, Lin CS, Leung CL, Liem RK. Nervous-tissue-specific elimination of microtubule-actin crosslinking factor 1a results in multiple developmental defects in the mouse brain. Mol Cell Neurosci. 2010 May;44(1):1-14. PubMed PMID: 20170731. Pubmed Central PMCID: 2847646.

124. Mefford HC, Clauin S, Sharp AJ, Moller RS, Ullmann R, Kapur R, et al. Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. Am J Hum Genet. 2007 Nov;81(5):1057-69. PubMed PMID: 17924346. Pubmed Central PMCID: 2265663. Epub 2007/10/10. eng.

125. Liao BY, Zhang J. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6987-92. PubMed PMID: 18458337. Pubmed Central PMCID: 2383943.

126. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013 Jul;15(7):565-74. PubMed PMID: 23788249. Pubmed Central PMCID: 3727274.

127. Bernhardt BA, Soucier D, Hanson K, Savage MS, Jackson L, Wapner RJ. Women's experiences receiving abnormal prenatal chromosomal microarray testing results. Genet Med. 2013 Feb;15(2):139-45. PubMed PMID: 22955112. Pubmed Central PMCID: 3877835.

128. McGillivray G, Rosenfeld JA, McKinlay Gardner RJ, Gillam LH. Genetic counselling and ethical issues with chromosome microarray analysis in prenatal testing. Prenat Diagn. 2012 Apr;32(4):389-95. PubMed PMID: 22467169.

129. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998 Mar;62(3):676-89. PubMed PMID: 9497246. Pubmed Central PMCID: 1376944.

130. Yurkiewicz IR, Korf BR, Lehmann LS. Prenatal whole-genome sequencing--is the quest to know a fetus's future ethical? N Engl J Med. 2014 Jan 16;370(3):195-7. PubMed PMID: 24428465.

131. Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, Caswell R, et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet. 2013 Nov 10. PubMed PMID: 24212882.

132. Dathe K, Kjaer KW, Brehm A, Meinecke P, Nurnberg P, Neto JC, et al. Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type A2. Am J Hum Genet. 2009 Apr;84(4):483-92. PubMed PMID: 19327734. Pubmed Central PMCID: 2667973.

133. Biesecker LG, Shianna KV, Mullikin JC. Exome sequencing: the expert view. Genome Biol. 2011;12(9):128. PubMed PMID: 21920051. Pubmed Central PMCID: 3308041.

134. Press release: DNA tests to revolutionise fight against cancer and help 100,000 NHS patients 2012. Available from: https://www.gov.uk/government/news/dna-tests-to-revolutionise-fight-against-cancer-and-help-100000-nhs-patients.

135. van Bokhoven H. Genetic and epigenetic networks in intellectual disabilities. Annu Rev Genet. 2011;45:81-104. PubMed PMID: 21910631.

136. Bryson SE, Bradley EA, Thompson A, Wainwright A. Prevalence of autism among adolescents with intellectual disabilities. Can J Psychiatry. 2008 Jul;53(7):449-59. PubMed PMID: 18674403.

137. Kaufman L, Ayub M, Vincent JB. The genetic basis of non-syndromic intellectual disability: a review. J Neurodev Disord. 2010 Dec;2(4):182-209. PubMed PMID: 21124998. Pubmed Central PMCID: 2974911.

138. Leonard H, Wen X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. Mental retardation and developmental disabilities research reviews. 2002;8(3):117-34. PubMed PMID: 12216056.

139. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, Saxena S. Prevalence of intellectual disability: a meta-analysis of population-based studies. Res Dev Disabil. 2011 Mar-Apr;32(2):419-36. PubMed PMID: 21236634.

140. Baird PA, Sadovnick AD. Mental retardation in over half-a-million consecutive livebirths: an epidemiological study. Am J Ment Defic. 1985 Jan;89(4):323-30. PubMed PMID: 3976730.

141. Centers for Disease C, Prevention. Economic costs associated with mental retardation, cerebral palsy, hearing loss, and vision impairment--United States, 2003. MMWR Morb Mortal Wkly Rep. 2004 Jan 30;53(3):57-9. PubMed PMID: 14749614.

142. Waber DP, Bryce CP, Girard JM, Zichlin M, Fitzmaurice GM, Galler JR. Impaired IQ and academic skills in adults who experienced moderate to severe infantile malnutrition: a 40-year study. Nutr Neurosci. 2014 Feb;17(2):58-64. PubMed PMID: 23484464. Pubmed Central PMCID: 3796166.

143. Niccols A. Fetal alcohol syndrome and the developing socio-emotional brain. Brain Cogn. 2007 Oct;65(1):135-42. PubMed PMID: 17669569.

144. Freij BJ, South MA, Sever JL. Maternal rubella and the congenital rubella syndrome. Clin Perinatol. 1988 Jun;15(2):247-57. PubMed PMID: 3288422.

145. Solon O, Riddell TJ, Quimbo SA, Butrick E, Aylward GP, Lou Bacate M, et al. Associations between cognitive function, blood lead concentration, and nutrition among children in the central Philippines. J Pediatr. 2008 Feb;152(2):237-43. PubMed PMID: 18206696.

146. Seidman LJ, Buka SL, Goldstein JM, Horton NJ, Rieder RO, Tsuang MT. The relationship of prenatal and perinatal complications to cognitive functioning at age 7 in the New England Cohorts of the National Collaborative Perinatal Project. Schizophr Bull. 2000;26(2):309-21. PubMed PMID: 10885633.

147. L. P. A clinical and genetic study of 1280 cases of mental defect. London: HMSO. 1938;229.

148. E. S. Mental Deficiency. The Eugenics Review. 1938;30(3):208-9.

149. Haldane JBS. A Clinical and Genetic Study of 1280 Cases of Mental Defect. Nature. 1938;141:575-6.

150. Deary IJ, Johnson W, Houlihan LM. Genetic foundations of human intelligence. Hum Genet. 2009 Jul;126(1):215-32. PubMed PMID: 19294424.

151. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013 Jun 21;340(6139):1467-71. PubMed PMID: 23722424. Pubmed Central PMCID: 3751588.

152. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A. 2006 Oct 1;140(19):2063-74. PubMed PMID: 16917849.

153. Lejeune J, Turpin R, Gautier M. [Mongolism; a chromosomal disease (trisomy)]. Bull Acad Natl Med. 1959 Apr 7-14;143(11-12):256-65. PubMed PMID: 13662687. Le mongolisme, maladie chromosomique. (trisomie).

154. Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, et al. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet. 2006 Sep;38(9):1032-7. PubMed PMID: 16906163.

155. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011 Aug 14. PubMed PMID: 21841781. Eng.

156. Talkowski ME, Mullegama SV, Rosenfeld JA, van Bon BW, Shen Y, Repnikova EA, et al. Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. Am J Hum Genet. 2011 Oct 7;89(4):551-63. PubMed PMID: 21981781. Pubmed Central PMCID: 3188839.

157. Zollino M, Orteschi D, Murdolo M, Lattante S, Battaglia D, Stefanini C, et al. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. Nat Genet. 2012 Jun;44(6):636-8. PubMed PMID: 22544367.

158. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet. 2010 Mar;42(3):203-9. PubMed PMID: 20154674. Pubmed Central PMCID: 2847896.

159. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint

cluster region exhibiting length variation in fragile X syndrome. Cell. 1991 May 31;65(5):905-14. PubMed PMID: 1710175.

160. Coffee B, Keith K, Albizua I, Malone T, Mowrey J, Sherman SL, et al. Incidence of fragile X syndrome by newborn screening for methylated FMR1 DNA. Am J Hum Genet. 2009 Oct;85(4):503-14. PubMed PMID: 19804849. Pubmed Central PMCID: 2756550.

161. Ausio J, Paz AM, Esteller M. MeCP2: the long trip from a chromatin protein to neurological disorders. Trends Mol Med. 2014 Apr 21. PubMed PMID: 24766768.

162. Ropers HH. Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet. 2010 Sep 22;11:161-87. PubMed PMID: 20822471.

163. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature. 2011 Oct 6;478(7367):57-63. PubMed PMID: 21937992. Epub 2011/09/23. eng.

164. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. Nat Genet. 2010 Dec;42(12):1109-12. PubMed PMID: 21076407.

165. Salvador-Carulla L, Bertelli M. 'Mental retardation' or 'intellectual disability': time for a conceptual change. Psychopathology. 2008;41(1):10-6. PubMed PMID: 17952016.

166. Saitsu H, Kato M, Mizuguchi T, Hamada K, Osaka H, Tohyama J, et al. De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. Nat Genet. 2008 Jun;40(6):782-8. PubMed PMID: 18469812.

167. Zhang Y, Luan Z, Liu A, Hu G. The scaffolding protein CASK mediates the interaction between rabphilin3a and beta-neurexins. FEBS Lett. 2001 May 25;497(2-3):99-102. PubMed PMID: 11377421.

168. Carrie A, Jun L, Bienvenu T, Vinet MC, McDonell N, Couvert P, et al. A new member of the IL-1 receptor family highly expressed in hippocampus and involved in X-linked mental retardation. Nat Genet. 1999 Sep;23(1):25-31. PubMed PMID: 10471494.

169. Wu Y, Arai AC, Rumbaugh G, Srivastava AK, Turner G, Hayashi T, et al. Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated

with moderate cognitive impairment in humans. Proc Natl Acad Sci U S A. 2007 Nov 13;104(46):18163-8. PubMed PMID: 17989220. Pubmed Central PMCID: 2084314.

170. Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, Pellerin S, et al. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N Engl J Med. 2009 Feb 5;360(6):599-605. PubMed PMID: 19196676. Pubmed Central PMCID: 2925262.

171. Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, et al. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nat Genet. 2010 Jun;42(6):489-91. PubMed PMID: 20473310.

172. Kleefstra T, Smidt M, Banning MJ, Oudakker AR, Van Esch H, de Brouwer AP, et al. Disruption of the gene Euchromatin Histone Methyl Transferase1 (Eu-HMTase1) is associated with the 9q34 subtelomeric deletion syndrome. J Med Genet. 2005 Apr;42(4):299-306. PubMed PMID: 15805155. Pubmed Central PMCID: 1736026.

173. Hoyer J, Ekici AB, Endele S, Popp B, Zweier C, Wiesener A, et al. Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. Am J Hum Genet. 2012 Mar 9;90(3):565-72. PubMed PMID: 22405089. Pubmed Central PMCID: 3309205. Epub 2012/03/13. eng.

174. Zalfa F, Eleuteri B, Dickson KS, Mercaldo V, De Rubeis S, di Penta A, et al. A new function for the fragile X mental retardation protein in regulation of PSD-95 mRNA stability. Nat Neurosci. 2007 May;10(5):578-87. PubMed PMID: 17417632. Pubmed Central PMCID: 2804293.

175. Paine RS. The variability in manifestations of untreated patients with phenylketonuria (phenylpyruvic aciduria). Pediatrics. 1957 Aug;20(2):290-302. PubMed PMID: 13452670.

176. de Lonlay P, Seta N, Barrot S, Chabrol B, Drouin V, Gabriel BM, et al. A broad spectrum of clinical presentations in congenital disorders of glycosylation I: a series of 26 cases. J Med Genet. 2001 Jan;38(1):14-9. PubMed PMID: 11134235. Pubmed Central PMCID: 1734729.

177. Fiala JC, Spacek J, Harris KM. Dendritic spine pathology: cause or consequence of neurological disorders? Brain Res Brain Res Rev. 2002 Jun;39(1):29-54. PubMed PMID: 12086707.

178. Zanni G, Saillour Y, Nagara M, Billuart P, Castelnau L, Moraine C, et al. Oligophrenin 1 mutations frequently cause X-linked mental retardation with cerebellar hypoplasia. Neurology. 2005 Nov 8;65(9):1364-9. PubMed PMID: 16221952.

179. Lebel RR, May M, Pouls S, Lubs HA, Stevenson RE, Schwartz CE. Non-syndromic X-linked mental retardation associated with a missense mutation (P312L) in the FGD1 gene. Clin Genet. 2002 Feb;61(2):139-45. PubMed PMID: 11940089.

180. Kerr B, Delrue MA, Sigaudy S, Perveen R, Marche M, Burgelin I, et al. Genotype-phenotype correlation in Costello syndrome: HRAS mutation analysis in 43 cases. J Med Genet. 2006 May;43(5):401-5. PubMed PMID: 16443854. Pubmed Central PMCID: 2564514.

181. San Martin A, Pagani MR. Understanding intellectual disability through RASopathies. J Physiol Paris. 2014 May 21. PubMed PMID: 24859216.

182. Barrientos RM, O'Reilly RC, Rudy JW. Memory for context is impaired by injecting anisomycin into dorsal hippocampus following context exploration. Behav Brain Res. 2002 Aug 21;134(1-2):299-306. PubMed PMID: 12191817.

183. Ohno H, Shinoda K, Ohyama K, Sharp LZ, Kajimura S. EHMT1 controls brown adipose cell fate and thermogenesis through the PRDM16 complex. Nature. 2013 Dec 5;504(7478):163-7. PubMed PMID: 24196706. Pubmed Central PMCID: 3855638.

184. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014 Jul 3;95(1):5-23. PubMed PMID: 24995866. Pubmed Central PMCID: 4085641.

185. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3;615(1-2):28-56. PubMed PMID: 17101154.

186. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004 Aug 6;305(5685):869-72. PubMed PMID: 15297675.

187. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics. 2009 Feb;5(2):e1000384. PubMed PMID: 19214210. Pubmed Central PMCID: 2633048.

188. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011 Jul 15;89(1):82-93. PubMed PMID: 21737059. Pubmed Central PMCID: 3135811.

189. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003 Feb 15;361(9357):598-604. PubMed PMID: 12598158.

190. Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS genetics. 2013 Apr;9(4):e1003443. PubMed PMID: 23593035. Pubmed Central PMCID: 3623759.

191. Grozeva D, Carss K, Spasic-Boskovic O, Parker MJ, Archer H, Firth HV, et al. De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. Am J Hum Genet. 2014 Apr 3;94(4):618-24. PubMed PMID: 24680889. Pubmed Central PMCID: 3980521.

192. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. PubMed PMID: 20644199. Pubmed Central PMCID: 2928508. Epub 2010/07/21. eng.

193. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010 Aug 15;26(16):2069-70. PubMed PMID: 20562413. Pubmed Central PMCID: 2916720. Epub 2010/06/22. eng.

194. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012 Dec 15;28(24):3326-8. PubMed PMID: 23060615. Pubmed Central PMCID: 3519454.

195. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2;467(7311):52-8. PubMed PMID: 20811451. Pubmed Central PMCID: 3173859.

196. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum

Genet. 2011 Apr 8;88(4):440-9. PubMed PMID: 21457909. Pubmed Central PMCID: 3071923.

197. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012 Apr 26;74(2):285-99. PubMed PMID: 22542183. Pubmed Central PMCID: 3619976.

198. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012 May 10;485(7397):242-5. PubMed PMID: 22495311. Pubmed Central PMCID: 3613847.

199. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014 May 1;94(5):677-94. PubMed PMID: 24768552. Pubmed Central PMCID: 4067558.

200. Kellogg G, Sum J, Wallerstein R. Deletion of 3p25.3 in a patient with intellectual disability and dysmorphic features with further definition of a critical region. Am J Med Genet A. 2013 Jun;161(6):1405-8. PubMed PMID: 23613140.

201. Verjaal M, De Nef MB. A patient with a partial deletion of the short arm of chromosome 3. Am J Dis Child. 1978 Jan;132(1):43-5. PubMed PMID: 623063.

202. Gunnarsson C, Foyn Bruun C. Molecular characterization and clinical features of a patient with an interstitial deletion of 3p25.3-p26.1. Am J Med Genet A. 2010 Dec;152A(12):3110-4. PubMed PMID: 21082655.

203. Riess A, Grasshoff U, Schaferhoff K, Bonin M, Riess O, Horber V, et al. Interstitial 3p25.3-p26.1 deletion in a patient with intellectual disability. Am J Med Genet A. 2012 Oct;158A(10):2587-90. PubMed PMID: 22965684.

204. Peltekova IT, Macdonald A, Armour CM. Microdeletion on 3p25 in a patient with features of 3p deletion syndrome. Am J Med Genet A. 2012 Oct;158A(10):2583-6. PubMed PMID: 22903836.

205. UniProt C. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D43-7. PubMed PMID: 23161681. Pubmed Central PMCID: 3531094.

206. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, et al. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science. 2002 Nov 1;298(5595):1039-43. PubMed PMID: 12351676.

207. Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, et al. Role of histone H3 lysine 27 methylation in X inactivation. Science. 2003 Apr 4;300(5616):131-5. PubMed PMID: 12649488.

208. Kamminga LM, Bystrykh LV, de Boer A, Houwer S, Douma J, Weersing E, et al. The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. Blood. 2006 Mar 1;107(5):2170-9. PubMed PMID: 16293602. Pubmed Central PMCID: 1895717.

209. Weaver DD, Graham CB, Thomas IT, Smith DW. A new overgrowth syndrome with accelerated skeletal maturation, unusual facies, and camptodactyly. J Pediatr. 1974 Apr;84(4):547-52. PubMed PMID: 4366187.

210. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010 Sep;42(9):790-3. PubMed PMID: 20711175. Pubmed Central PMCID: 2930028.

211. Zechner U, Wilda M, Kehrer-Sawatzki H, Vogel W, Fundele R, Hameister H. A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? Trends Genet. 2001 Dec;17(12):697-701. PubMed PMID: 11718922.

212. Nguyen DK, Disteche CM. High expression of the mammalian X chromosome in brain. Brain Res. 2006 Dec 18;1126(1):46-9. PubMed PMID: 16978591.

213. Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, et al. Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. Nat Genet. 2011 Dec;43(12):1179-85. PubMed PMID: 22019781. Pubmed Central PMCID: 3576853.

214. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet. 2010 Oct;42(10):830-1. PubMed PMID: 20802480.

215. Schaffner SF. The X chromosome in population genetics. Nature reviews Genetics. 2004 Jan;5(1):43-51. PubMed PMID: 14708015.

216. Beltran-Valero de Bernabe D, Currier S, Steinbrecher A, Celli J, van Beusekom E, van der Zwaag B, et al. Mutations in the O-mannosyltransferase gene POMT1 give rise to the severe neuronal migration disorder Walker-Warburg syndrome. Am J Hum

Genet. 2002 Nov;71(5):1033-43. PubMed PMID: 12369018. Pubmed Central PMCID: 419999. Epub 2002/10/09. eng.

217. Yoshida A, Kobayashi K, Manya H, Taniguchi K, Kano H, Mizuno M, et al. Muscular dystrophy and neuronal migration disorder caused by mutations in a glycosyltransferase, POMGnT1. Dev Cell. 2001 Nov;1(5):717-24. PubMed PMID: 11709191. Epub 2001/11/16. eng.

218. van Reeuwijk J, Maugenre S, van den Elzen C, Verrips A, Bertini E, Muntoni F, et al. The expanding phenotype of POMT1 mutations: from Walker-Warburg syndrome to congenital muscular dystrophy, microcephaly, and mental retardation. Hum Mutat. 2006 May;27(5):453-9. PubMed PMID: 16575835.

219. Godfrey C, Clement E, Mein R, Brockington M, Smith J, Talim B, et al. Refining genotype phenotype correlations in muscular dystrophies with defective glycosylation of dystroglycan. Brain : a journal of neurology. 2007 Oct;130(Pt 10):2725-35. PubMed PMID: 17878207. Epub 2007/09/20. eng.

220. Ibraghimov-Beskrovnaya O, Ervasti JM, Leveille CJ, Slaughter CA, Sernett SW, Campbell KP. Primary structure of dystrophin-associated glycoproteins linking dystrophin to the extracellular matrix. Nature. 1992 Feb 20;355(6362):696-702. PubMed PMID: 1741056. Epub 1992/02/20. eng.

221. van Reeuwijk J, Janssen M, van den Elzen C, Beltran-Valero de Bernabe D, Sabatelli P, Merlini L, et al. POMT2 mutations cause alpha-dystroglycan hypoglycosylation and Walker-Warburg syndrome. J Med Genet. 2005 Dec;42(12):907-12. PubMed PMID: 15894594. Pubmed Central PMCID: 1735967. Epub 2005/05/17. eng.

222. Wright KM, Lyon KA, Leung H, Leahy DJ, Ma L, Ginty DD. Dystroglycan organizes axon guidance cue localization and axonal pathfinding. Neuron. 2012 Dec 6;76(5):931-44. PubMed PMID: 23217742. Pubmed Central PMCID: 3526105. Epub 2012/12/12. eng.

223. Wells L. The o-mannosylation pathway: glycosyltransferases and proteins implicated in congenital muscular dystrophy. J Biol Chem. 2013 Mar 8;288(10):6930-5. PubMed PMID: 23329833. Pubmed Central PMCID: 3591603.

224. Stalnaker SH, Hashmi S, Lim JM, Aoki K, Porterfield M, Gutierrez-Sanchez G, et al. Site mapping and characterization of O-glycan structures on alpha-dystroglycan

isolated from rabbit skeletal muscle. J Biol Chem. 2010 Aug 6;285(32):24882-91. PubMed PMID: 20507986. Pubmed Central PMCID: 2915724. Epub 2010/05/29. eng.

225. Tran DT, Lim JM, Liu M, Stalnaker SH, Wells L, Ten Hagen KG, et al. Glycosylation of alpha-dystroglycan: O-mannosylation influences the subsequent addition of GalNAc by UDP-GalNAc polypeptide N-acetylgalactosaminyltransferases. J Biol Chem. 2012 Jun 15;287(25):20967-74. PubMed PMID: 22549772. Pubmed Central PMCID: 3375520.

226. Michele DE, Barresi R, Kanagawa M, Saito F, Cohn RD, Satz JS, et al. Post-translational disruption of dystroglycan-ligand interactions in congenital muscular dystrophies. Nature. 2002 Jul 25;418(6896):417-22. PubMed PMID: 12140558. eng.

227. Yoshida-Moriguchi T, Yu L, Stalnaker SH, Davis S, Kunz S, Madson M, et al. O-mannosyl phosphorylation of alpha-dystroglycan is required for laminin binding. Science. 2010 Jan 1;327(5961):88-92. PubMed PMID: 20044576. Pubmed Central PMCID: 2978000. Epub 2010/01/02. eng.

228. Frost AR, Bohm SV, Sewduth RN, Josifova D, Ogilvie CM, Izatt L, et al. Heterozygous deletion of a 2-Mb region including the dystroglycan gene in a patient with mild myopathy, facial hypotonia, oral-motor dyspraxia and white matter abnormalities. Eur J Hum Genet. 2010 Jul;18(7):852-5. PubMed PMID: 20234391. Pubmed Central PMCID: 2987357.

229. Hara Y, Balci-Hayta B, Yoshida-Moriguchi T, Kanagawa M, Beltran-Valero de Bernabe D, Gundesli H, et al. A dystroglycan mutation associated with limb-girdle muscular dystrophy. N Engl J Med. 2011 Mar 10;364(10):939-46. PubMed PMID: 21388311. Pubmed Central PMCID: 3071687. Epub 2011/03/11. eng.

230. Geis T, Marquard K, Rodl T, Reihle C, Schirmer S, von Kalle T, et al. Homozygous dystroglycan mutation associated with a novel muscle-eye-brain disease-like phenotype with multicystic leucodystrophy. Neurogenetics. 2013 Nov;14(3-4):205-13. PubMed PMID: 24052401.

231. Maeda Y, Tanaka S, Hino J, Kangawa K, Kinoshita T. Human dolichol-phosphate-mannose synthase consists of three subunits, DPM1, DPM2 and DPM3. EMBO J. 2000 Jun 1;19(11):2475-82. PubMed PMID: 10835346. Pubmed Central PMCID: 212771.

232. Ashida H, Maeda Y, Kinoshita T. DPM1, the catalytic subunit of dolichol-phosphate mannose synthase, is tethered to and stabilized on the endoplasmic

reticulum membrane by DPM3. J Biol Chem. 2006 Jan 13;281(2):896-904. PubMed PMID: 16280320.

233. Kim S, Westphal V, Srikrishna G, Mehta DP, Peterson S, Filiano J, et al. Dolichol phosphate mannose synthase (DPM1) mutations define congenital disorder of glycosylation Ie (CDG-Ie). J Clin Invest. 2000 Jan;105(2):191-8. PubMed PMID: 10642597. Pubmed Central PMCID: 377427.

234. Lefeber DJ, Schonberger J, Morava E, Guillard M, Huyben KM, Verrijp K, et al. Deficiency of Dol-P-Man synthase subunit DPM3 bridges the congenital disorders of glycosylation with the dystroglycanopathies. Am J Hum Genet. 2009 Jul;85(1):76-86. PubMed PMID: 19576565. eng.

235. Barone R, Aiello C, Race V, Morava E, Foulquier F, Riemersma M, et al. DPM2-CDG: A muscular dystrophy-dystroglycanopathy syndrome with severe epilepsy. Ann Neurol. 2012 Oct;72(4):550-8. PubMed PMID: 23109149. Epub 2012/10/31. eng.

236. Yang AC, Ng BG, Moore SA, Rush J, Waechter CJ, Raymond KM, et al. Congenital disorder of glycosylation due to DPM1 mutations presenting with dystroglycanopathy-type congenital muscular dystrophy. Mol Genet Metab. 2013 Nov;110(3):345-51. PubMed PMID: 23856421. Pubmed Central PMCID: 3800268.

237. Lefeber DJ, de Brouwer AP, Morava E, Riemersma M, Schuurs-Hoeijmakers JH, Absmanner B, et al. Autosomal recessive dilated cardiomyopathy due to DOLK mutations results from abnormal dystroglycan O-mannosylation. PLoS genetics. 2011 Dec;7(12):e1002427. PubMed PMID: 22242004. Pubmed Central PMCID: 3248466. Epub 2012/01/14. eng.

238. Manya H, Chiba A, Yoshida A, Wang X, Chiba Y, Jigami Y, et al. Demonstration of mammalian protein O-mannosyltransferase activity: coexpression of POMT1 and POMT2 required for enzymatic activity. Proc Natl Acad Sci U S A. 2004 Jan 13;101(2):500-5. PubMed PMID: 14699049. Pubmed Central PMCID: 327176.

239. Akasaka-Manya K, Manya H, Nakajima A, Kawakita M, Endo T. Physical and functional association of human protein O-mannosyltransferases 1 and 2. J Biol Chem. 2006 Jul 14;281(28):19339-45. PubMed PMID: 16698797.

240. Akasaka-Manya K, Manya H, Hayashi M, Endo T. Different roles of the two components of human protein O-mannosyltransferase, POMT1 and POMT2. Biochem Biophys Res Commun. 2011 Aug 12;411(4):721-5. PubMed PMID: 21782786.

241. Hehr U, Uyanik G, Gross C, Walter MC, Bohring A, Cohen M, et al. Novel POMGnT1 mutations define broader phenotypic spectrum of muscle-eye-brain disease. Neurogenetics. 2007 Nov;8(4):279-88. PubMed PMID: 17906881.

242. Raducu M, Baets J, Fano O, Van Coster R, Cruces J. Promoter alteration causes transcriptional repression of the POMGNT1 gene in limb-girdle muscular dystrophy type 2O. Eur J Hum Genet. 2012 Sep;20(9):945-52. PubMed PMID: 22419172. Pubmed Central PMCID: 3421125.

243. Vuillaumier-Barrot S, Bouchet-Seraphin C, Chelbi M, Eude-Caye A, Charluteau E, Besson C, et al. Intragenic rearrangements in LARGE and POMGNT1 genes in severe dystroglycanopathies. Neuromuscul Disord. 2011 Nov;21(11):782-90. PubMed PMID: 21727005.

244. Balci B, Uyanik G, Dincer P, Gross C, Willer T, Talim B, et al. An autosomal recessive limb girdle muscular dystrophy (LGMD2) with mild mental retardation is allelic to Walker-Warburg syndrome (WWS) caused by a mutation in the POMT1 gene. Neuromuscul Disord. 2005 Apr;15(4):271-5. PubMed PMID: 15792865.

245. Bello L, Melacini P, Pezzani R, D'Amico A, Piva L, Leonardi E, et al. Cardiomyopathy in patients with POMT1-related congenital and limb-girdle muscular dystrophy. Eur J Hum Genet. 2012 Dec;20(12):1234-9. PubMed PMID: 22549409. Pubmed Central PMCID: 3499746.

246. Yanagisawa A, Bouchet C, Quijano-Roy S, Vuillaumier-Barrot S, Clarke N, Odent S, et al. POMT2 intragenic deletions and splicing abnormalities causing congenital muscular dystrophy with mental retardation. Eur J Med Genet. 2009 Jul-Aug;52(4):201-6. PubMed PMID: 19138766.

247. Akasaka-Manya K, Manya H, Endo T. Mutations of the POMT1 gene found in patients with Walker-Warburg syndrome lead to a defect of protein O-mannosylation. Biochem Biophys Res Commun. 2004 Dec 3;325(1):75-9. PubMed PMID: 15522202.

248. Kanagawa M, Saito F, Kunz S, Yoshida-Moriguchi T, Barresi R, Kobayashi YM, et al. Molecular recognition by LARGE is essential for expression of functional dystroglycan. Cell. 2004 Jun 25;117(7):953-64. PubMed PMID: 15210115.

249. Inamori K, Yoshida-Moriguchi T, Hara Y, Anderson ME, Yu L, Campbell KP. Dystroglycan function requires xylosyl- and glucuronyltransferase activities of LARGE. Science. 2012 Jan 6;335(6064):93-6. PubMed PMID: 22223806. Epub 2012/01/10. eng.

250. Longman C, Brockington M, Torelli S, Jimenez-Mallebrera C, Kennedy C, Khalil N, et al. Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. Hum Mol Genet. 2003 Nov 1;12(21):2853-61. PubMed PMID: 12966029. Epub 2003/09/11. eng.

251. Brockington M, Torelli S, Prandini P, Boito C, Dolatshad NF, Longman C, et al. Localization and functional analysis of the LARGE family of glycosyltransferases: significance for muscular dystrophy. Hum Mol Genet. 2005 Mar 1;14(5):657-65. PubMed PMID: 15661757.

252. Bao X, Kobayashi M, Hatakeyama S, Angata K, Gullberg D, Nakayama J, et al. Tumor suppressor function of laminin-binding alpha-dystroglycan requires a distinct beta3-N-acetylglucosaminyltransferase. Proc Natl Acad Sci U S A. 2009 Jul 21;106(29):12109-14. PubMed PMID: 19587235. eng.

253. Lee PL, Kohler JJ, Pfeffer SR. Association of beta-1,3-N-acetylglucosaminyltransferase 1 and beta-1,4-galactosyltransferase 1, trans-Golgi enzymes involved in coupled poly-N-acetyllactosamine synthesis. Glycobiology. 2009 Jun;19(6):655-64. PubMed PMID: 19261593. Pubmed Central PMCID: 2682609.

254. Manzini MC, Tambunan DE, Hill RS, Yu TW, Maynard TM, Heinzen EL, et al. Exome Sequencing and Functional Validation in Zebrafish Identify GTDC2 Mutations as a Cause of Walker-Warburg Syndrome. Am J Hum Genet. 2012 Sep 7;91(3):541-7. PubMed PMID: 22958903. Epub 2012/09/11. eng.

255. Ogawa M, Nakamura N, Nakayama Y, Kurosaka A, Manya H, Kanagawa M, et al. GTDC2 modifies O-mannosylated alpha-dystroglycan in the endoplasmic reticulum to generate N-acetyl glucosamine epitopes reactive with CTD110.6 antibody. Biochem Biophys Res Commun. 2013 Oct 11;440(1):88-93. PubMed PMID: 24041696.

256. Jae LT, Raaben M, Riemersma M, van Beusekom E, Blomen VA, Velds A, et al. Deciphering the Glycosylome of Dystroglycanopathies Using Haploid Screens for Lassa Virus Entry. Science. 2013 Mar 21. PubMed PMID: 23519211. Epub 2013/03/23. Eng.

257. von Renesse A, Petkova MV, Lutzkendorf S, Heinemeyer J, Gill E, Hubner C, et al. POMK mutation in a family with congenital muscular dystrophy with merosin deficiency, hypomyelination, mild hearing deficit and intellectual disability. J Med Genet. 2014 Feb 20. PubMed PMID: 24556084.

258. Yoshida-Moriguchi T, Willer T, Anderson ME, Venzke D, Whyte T, Muntoni F, et al. SGK196 is a glycosylation-specific O-mannose kinase required for dystroglycan function. Science. 2013 Aug 23;341(6148):896-9. PubMed PMID: 23929950. Pubmed Central PMCID: 3848040.

259. Toda T, Kobayashi K, Kondo-Iida E, Sasaki J, Nakamura Y. The Fukuyama congenital muscular dystrophy story. Neuromuscul Disord. 2000 Mar;10(3):153-9. PubMed PMID: 10734260.

260. Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature. 1998 Jul 23;394(6691):388-92. PubMed PMID: 9690476. Epub 1998/08/05. eng.

261. Watanabe M, Kobayashi K, Jin F, Park KS, Yamada T, Tokunaga K, et al. Founder SVA retrotransposal insertion in Fukuyama-type congenital muscular dystrophy and its origin in Japanese and Northeast Asian populations. Am J Med Genet A. 2005 Nov 1;138(4):344-8. PubMed PMID: 16222679.

262. Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu CC, Mori K, Oda T, et al. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature. 2011 Oct 6;478(7367):127-31. PubMed PMID: 21979053. Pubmed Central PMCID: 3412178.

263. Kondo-Iida E, Kobayashi K, Watanabe M, Sasaki J, Kumagai T, Koide H, et al. Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Hum Mol Genet. 1999 Nov;8(12):2303-9. PubMed PMID: 10545611.

264. Silan F, Yoshioka M, Kobayashi K, Simsek E, Tunc M, Alper M, et al. A new mutation of the fukutin gene in a non-Japanese patient. Ann Neurol. 2003 Mar;53(3):392-6. PubMed PMID: 12601708.

265. Cotarelo RP, Valero MC, Prados B, Pena A, Rodriguez L, Fano O, et al. Two new patients bearing mutations in the fukutin gene confirm the relevance of this gene in Walker-Warburg syndrome. Clin Genet. 2008 Feb;73(2):139-45. PubMed PMID: 18177472.

266. Godfrey C, Escolar D, Brockington M, Clement EM, Mein R, Jimenez-Mallebrera C, et al. Fukutin gene mutations in steroid-responsive limb girdle muscular dystrophy. Ann Neurol. 2006 Nov;60(5):603-10. PubMed PMID: 17044012.

267. Yis U, Uyanik G, Heck PB, Smitka M, Nobel H, Ebinger F, et al. Fukutin mutations in non-Japanese patients with congenital muscular dystrophy: less severe mutations predominate in patients with a non-Walker-Warburg phenotype. Neuromuscul Disord. 2011 Jan;21(1):20-30. PubMed PMID: 20961758.

268. Hayashi YK, Ogawa M, Tagawa K, Noguchi S, Ishihara T, Nonaka I, et al. Selective deficiency of alpha-dystroglycan in Fukuyama-type congenital muscular dystrophy. Neurology. 2001 Jul 10;57(1):115-21. PubMed PMID: 11445638.

269. Brockington M, Blake DJ, Prandini P, Brown SC, Torelli S, Benson MA, et al. Mutations in the fukutin-related protein gene (FKRP) cause a form of congenital muscular dystrophy with secondary laminin alpha2 deficiency and abnormal glycosylation of alpha-dystroglycan. Am J Hum Genet. 2001 Dec;69(6):1198-209. PubMed PMID: 11592034. Pubmed Central PMCID: 1235559. Epub 2001/10/10. eng.

270. Brockington M, Yuva Y, Prandini P, Brown SC, Torelli S, Benson MA, et al. Mutations in the fukutin-related protein gene (FKRP) identify limb girdle muscular dystrophy 2I as a milder allelic variant of congenital muscular dystrophy MDC1C. Hum Mol Genet. 2001 Dec 1;10(25):2851-9. PubMed PMID: 11741828.

271. Mercuri E, Brockington M, Straub V, Quijano-Roy S, Yuva Y, Herrmann R, et al. Phenotypic spectrum associated with mutations in the fukutin-related protein gene. Ann Neurol. 2003 Apr;53(4):537-42. PubMed PMID: 12666124.

272. Bourteel H, Vermersch P, Cuisset JM, Maurage CA, Laforet P, Richard P, et al. Clinical and mutational spectrum of limb-girdle muscular dystrophy type 2I in 11 French patients. J Neurol Neurosurg Psychiatry. 2009 Dec;80(12):1405-8. PubMed PMID: 19917824.

273. Saito Y, Mizuguchi M, Oka A, Takashima S. Fukutin protein is expressed in neurons of the normal developing human brain but is reduced in Fukuyama-type congenital muscular dystrophy brain. Ann Neurol. 2000 Jun;47(6):756-64. PubMed PMID: 10852541.

274. Yamamoto T, Kato Y, Karita M, Takeiri H, Muramatsu F, Kobayashi M, et al. Fukutin expression in glial cells and neurons: implication in the brain lesions of Fukuyama congenital muscular dystrophy. Acta Neuropathol. 2002 Sep;104(3):217-24. PubMed PMID: 12172906.

275. Matsumoto H, Noguchi S, Sugie K, Ogawa M, Murayama K, Hayashi YK, et al. Subcellular localization of fukutin and fukutin-related protein in muscle cells. J Biochem. 2004 Jun;135(6):709-12. PubMed PMID: 15213246.

276. Tachikawa M, Kanagawa M, Yu CC, Kobayashi K, Toda T. Mislocalization of fukutin protein by disease-causing missense mutations can be rescued with treatments directed at folding amelioration. J Biol Chem. 2012 Mar 9;287(11):8398-406. PubMed PMID: 22275357. Pubmed Central PMCID: 3318729.

277. Xiong H, Kobayashi K, Tachikawa M, Manya H, Takeda S, Chiyonobu T, et al. Molecular interaction between fukutin and POMGnT1 in the glycosylation pathway of alpha-dystroglycan. Biochem Biophys Res Commun. 2006 Dec 1;350(4):935-41. PubMed PMID: 17034757.

278. Beedle AM, Nienaber PM, Campbell KP. Fukutin-related protein associates with the sarcolemmal dystrophin-glycoprotein complex. J Biol Chem. 2007 Jun 8;282(23):16713-7. PubMed PMID: 17452335.

279. Esapa CT, McIlhinney RA, Blake DJ. Fukutin-related protein mutations that cause congenital muscular dystrophy result in ER-retention of the mutant protein in cultured cells. Hum Mol Genet. 2005 Jan 15;14(2):295-305. PubMed PMID: 15574464.

280. Dolatshad NF, Brockington M, Torelli S, Skordis L, Wever U, Wells DJ, et al. Mutated fukutin-related protein (FKRP) localises as wild type in differentiated muscle cells. Exp Cell Res. 2005 Oct 1;309(2):370-8. PubMed PMID: 16055117.

281. Willer T, Lee H, Lommel M, Yoshida-Moriguchi T, de Bernabe DB, Venzke D, et al. ISPD loss-of-function mutations disrupt dystroglycan O-mannosylation and cause Walker-Warburg syndrome. Nat Genet. 2012;44(5):575-80. PubMed PMID: 22522420. Pubmed Central PMCID: 3371168. Epub 2012/04/24. eng.

282. Roscioli T, Kamsteeg EJ, Buysse K, Maystadt I, van Reeuwijk J, van den Elzen C, et al. Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of alpha-dystroglycan. Nat Genet. 2012;44(5):581-5. PubMed PMID: 22522421. Epub 2012/04/24. eng.

283. Cirak S, Foley AR, Herrmann R, Willer T, Yau S, Stevens E, et al. ISPD gene mutations are a common cause of congenital and limb-girdle muscular dystrophies. Brain : a journal of neurology. 2013 Jan 3. PubMed PMID: 23288328. Epub 2013/01/05. Eng.

284. Vuillaumier-Barrot S, Bouchet-Seraphin C, Chelbi M, Devisme L, Quentin S, Gazal S, et al. Identification of Mutations in TMEM5 and ISPD as a Cause of Severe Cobblestone Lissencephaly. Am J Hum Genet. 2012 Dec 7;91(6):1135-43. PubMed PMID: 23217329. Pubmed Central PMCID: 3516603. Epub 2012/12/12. eng.

285. Peat RA, Smith JM, Compton AG, Baker NL, Pace RA, Burkin DJ, et al. Diagnosis and etiology of congenital muscular dystrophy. Neurology. 2008 Jul 29;71(5):312-21. PubMed PMID: 18160674.

286. Messina S, Mora M, Pegoraro E, Pini A, Mongini T, D'Amico A, et al. POMT1 and POMT2 mutations in CMD patients: a multicentric Italian study. Neuromuscul Disord. 2008 Jul;18(7):565-71. PubMed PMID: 18513969.

287. Mercuri E, Messina S, Bruno C, Mora M, Pegoraro E, Comi GP, et al. Congenital muscular dystrophies with defective glycosylation of dystroglycan: a population study. Neurology. 2009 May 26;72(21):1802-9. PubMed PMID: 19299310. eng.

288. Vervoort VS, Holden KR, Ukadike KC, Collins JS, Saul RA, Srivastava AK. POMGnT1 gene alterations in a family with neurological abnormalities. Ann Neurol. 2004 Jul;56(1):143-8. PubMed PMID: 15236414.

289. Yanagisawa A, Bouchet C, Van den Bergh PY, Cuisset JM, Viollet L, Leturcq F, et al. New POMT2 mutations causing congenital muscular dystrophy: identification of a founder mutation. Neurology. 2007 Sep 18;69(12):1254-60. PubMed PMID: 17634419.

290. Clement E, Mercuri E, Godfrey C, Smith J, Robb S, Kinali M, et al. Brain involvement in muscular dystrophies with defective dystroglycan glycosylation. Ann Neurol. 2008 Nov;64(5):573-82. PubMed PMID: 19067344.

291. Van Reeuwijk J, Olderode-Berends MJ, Van den Elzen C, Brouwer OF, Roscioli T, Van Pampus MG, et al. A homozygous FKRP start codon mutation is associated with Walker-Warburg syndrome, the severe end of the clinical spectrum. Clin Genet. 2010 Sep;78(3):275-81. PubMed PMID: 20236121.

292. Jimenez-Mallebrera C, Torelli S, Feng L, Kim J, Godfrey C, Clement E, et al. A comparative study of alpha-dystroglycan glycosylation in dystroglycanopathies suggests that the hypoglycosylation of alpha-dystroglycan does not consistently correlate with clinical severity. Brain Pathol. 2009 Oct;19(4):596-611. PubMed PMID: 18691338. Pubmed Central PMCID: 2860390. Epub 2008/08/12. eng.

293. Saredi S, Ardissone A, Ruggieri A, Mottarelli E, Farina L, Rinaldi R, et al. Novel POMGNT1 point mutations and intragenic rearrangements associated with muscle-eye-brain disease. J Neurol Sci. 2012 Jul 15;318(1-2):45-50. PubMed PMID: 22554691. Pubmed Central PMCID: 3405532.

294. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013 Apr 25;496(7446):498-503. PubMed PMID: 23594743. Pubmed Central PMCID: 3703927.

295. Lin YY. Muscle diseases in the zebrafish. Neuromuscular disorders : NMD. 2012 May 28. PubMed PMID: 22647769. Epub 2012/06/01. Eng.

296. Parsons MJ, Campos I, Hirst EM, Stemple DL. Removal of dystroglycan causes severe muscular dystrophy in zebrafish embryos. Development. 2002 Jul;129(14):3505-12. PubMed PMID: 12091319. Epub 2002/07/02. eng.

297. Garrity DM, Childs S, Fishman MC. The heartstrings mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. Development. 2002 Oct;129(19):4635-45. PubMed PMID: 12223419.

298. Gerlai R. Social behavior of zebrafish: From synthetic images to biological mechanisms of shoaling. J Neurosci Methods. 2014 May 2. PubMed PMID: 24793400.

299. Kettleborough RN, Busch-Nentwich EM, Harvey SA, Dooley CM, de Bruijn E, van Eeden F, et al. A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature. 2013 Apr 25;496(7446):494-7. PubMed PMID: 23594742. Pubmed Central PMCID: 3743023.

300. Blackburn PR, Campbell JM, Clark KJ, Ekker SC. The CRISPR system--keeping zebrafish gene targeting fresh. Zebrafish. 2013 Mar;10(1):116-8. PubMed PMID: 23536990. Pubmed Central PMCID: 3629780.

301. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013 Sep;31(9):822-6. PubMed PMID: 23792628. Pubmed Central PMCID: 3773023.

302. Robu ME, Larson JD, Nasevicius A, Beiraghi S, Brenner C, Farber SA, et al. p53 activation by knockdown technologies. PLoS genetics. 2007 May 25;3(5):e78. PubMed PMID: 17530925. eng.

303. Gerety SS, Wilkinson DG. Morpholino artifacts provide pitfalls and reveal a novel role for pro-apoptotic genes in hindbrain boundary development. Dev Biol. 2011 Feb 15;350(2):279-89. PubMed PMID: 21145318. Pubmed Central PMCID: 3111810. Epub 2010/12/15. eng.

304. Moore CJ, Goh HT, Hewitt JE. Genes required for functional glycosylation of dystroglycan are conserved in zebrafish. Genomics. 2008 Sep;92(3):159-67. PubMed PMID: 18632251. Epub 2008/07/18. eng.

305. Kawahara G, Guyon JR, Nakamura Y, Kunkel LM. Zebrafish models for human FKRP muscular dystrophies. Hum Mol Genet. 2010 Feb 15;19(4):623-33. PubMed PMID: 19955119. Pubmed Central PMCID: 2807370. Epub 2009/12/04. eng.

306. Avsar-Ban E, Ishikawa H, Manya H, Watanabe M, Akiyama S, Miyake H, et al. Protein O-mannosylation is necessary for normal embryonic development in zebrafish. Glycobiology. 2010 Sep;20(9):1089-102. PubMed PMID: 20466645. Epub 2010/05/15. eng.

307. Gupta V, Kawahara G, Gundry SR, Chen AT, Lencer WI, Zhou Y, et al. The zebrafish dag1 mutant: a novel genetic model for dystroglycanopathies. Hum Mol Genet. 2011 May 1;20(9):1712-25. PubMed PMID: 21296866. Pubmed Central PMCID: 3071669.

308. Chiyonobu T, Sasaki J, Nagai Y, Takeda S, Funakoshi H, Nakamura T, et al. Effects of fukutin deficiency in the developing mouse brain. Neuromuscul Disord. 2005 Jun;15(6):416-26. PubMed PMID: 15907289.

309. Stalnaker SH, Aoki K, Lim JM, Porterfield M, Liu M, Satz JS, et al. Glycomic analyses of mouse models of congenital muscular dystrophy. J Biol Chem. 2011 Jun 17;286(24):21180-90. PubMed PMID: 21460210. Pubmed Central PMCID: 3122180.

310. Lin YY, White RJ, Torelli S, Cirak S, Muntoni F, Stemple DL. Zebrafish Fukutin family proteins link the unfolded protein response with dystroglycanopathies. Hum Mol Genet. 2011 May 1;20(9):1763-75. PubMed PMID: 21317159. eng.

311. Barresi R, Michele DE, Kanagawa M, Harper HA, Dovico SA, Satz JS, et al. LARGE can functionally bypass alpha-dystroglycan glycosylation defects in distinct congenital muscular dystrophies. Nat Med. 2004 Jul;10(7):696-703. PubMed PMID: 15184894. Epub 2004/06/09. eng.

312. Yu M, He Y, Wang K, Zhang P, Zhang S, Hu H. Adeno-associated viral-mediated LARGE gene therapy rescues the muscular dystrophic phenotype in mouse models of dystroglycanopathy. Hum Gene Ther. 2013 Mar;24(3):317-30. PubMed PMID: 23379513. Pubmed Central PMCID: 3609641.

313. Carss KJ, Stevens E, Foley AR, Cirak S, Riemersma M, Torelli S, et al. Mutations in GDP-mannose pyrophosphorylase B cause congenital and limb-girdle muscular dystrophies associated with hypoglycosylation of alpha-dystroglycan. Am J Hum Genet. 2013 Jul 11;93(1):29-41. PubMed PMID: 23768512. Pubmed Central PMCID: 3710768.

314. Stevens E, Carss KJ, Cirak S, Foley AR, Torelli S, Willer T, et al. Mutations in B3GALNT2 Cause Congenital Muscular Dystrophy and Hypoglycosylation of alpha-Dystroglycan. Am J Hum Genet. 2013 Mar 7;92(3):354-65. PubMed PMID: 23453667. Pubmed Central PMCID: 3591840. Epub 2013/03/05. eng.

315. Hiruma T, Togayachi A, Okamura K, Sato T, Kikuchi N, Kwon YD, et al. A novel human beta1,3-N-acetylgalactosaminyltransferase that synthesizes a unique carbohydrate structure, GalNAcbeta1-3GlcNAc. J Biol Chem. 2004 Apr 2;279(14):14087-95. PubMed PMID: 14724282. eng.

316. Harrison R, Hitchen PG, Panico M, Morris HR, Mekhaiel D, Pleass RJ, et al. Glycoproteomic characterization of recombinant mouse alpha-dystroglycan. Glycobiology. 2012 May;22(5):662-75. PubMed PMID: 22241827. Pubmed Central PMCID: 3311285.

317. Ning B, Elbein AD. Cloning, expression and characterization of the pig liver GDP-mannose pyrophosphorylase. Evidence that GDP-mannose and GDP-Glc pyrophosphorylases are different proteins. Eur J Biochem. 2000 Dec;267(23):6866-74. PubMed PMID: 11082198. eng.

318. Malicki J, Jo H, Wei X, Hsiung M, Pujic Z. Analysis of gene function in the zebrafish retina. Methods. 2002 Dec;28(4):427-38. PubMed PMID: 12507461. Epub 2003/01/01. eng.

319. Abramoff MD, Magalhaes, P.J., Ram, S.J. Image Processing with ImageJ. Biophotonics International. 2004;11(7):36-42.

320. Thisse B, Thisse, C. Fast Release Clones: A High Throughput Expression Analysis. ZFIN Direct Data Submission 2004. Available from: http://zfin.org.

321. Thisse B, Pflumio, S., Fürthauer, M., Loppin, B., Heyer, V., Degrave, A., Woehl, R., Lux, A., Steffan, T., Charbonnier, X.Q., Thisse, C. Expression of the zebrafish genome during embryogenesis (NIH R01 RR15402). ZFIN Direct Data Submission (http://zfinorg) 2001.

322. Bassett DI, Bryson-Richardson RJ, Daggett DF, Gautier P, Keenan DG, Currie PD. Dystrophin is required for the formation of stable muscle attachments in the zebrafish embryo. Development. 2003 Dec;130(23):5851-60. PubMed PMID: 14573513. eng.

323. Niederriter AR, Davis EE, Golzio C, Oh EC, Tsai IC, Katsanis N. In vivo modeling of the morbid human genome using Danio rerio. Journal of visualized experiments : JoVE. 2013 (78):e50338. PubMed PMID: 23995499. Pubmed Central PMCID: 3856313.

324. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012 May 17;485(7398):363-7. PubMed PMID: 22596160. Pubmed Central PMCID: 3366115. Epub 2012/05/19. eng.

325. Piepenburg O, Grimmer D, Williams PH, Smith JC. Activin redux: specification of mesodermal pattern in Xenopus by graded concentrations of endogenous activin B. Development. 2004 Oct;131(20):4977-86. PubMed PMID: 15371302.

326. Eisen JS, Smith JC. Controlling morpholino experiments: don't stop making antisense. Development. 2008 May;135(10):1735-43. PubMed PMID: 18403413.

327. Lisenbee CS, Karnik SK, Trelease RN. Overexpression and mislocalization of a tail-anchored GFP redefines the identity of peroxisomal ER. Traffic. 2003 Jul;4(7):491-501. PubMed PMID: 12795694.

328. Hedberg C, Oldfors A, Darin N. B3GALNT2 is a gene associated with congenital muscular dystrophy with brain malformations. Eur J Hum Genet. 2014 May;22(5):707-10. PubMed PMID: 24084573. Pubmed Central PMCID: 3992579.

329. Cao W, Henry MD, Borrow P, Yamada H, Elder JH, Ravkov EV, et al. Identification of alpha-dystroglycan as a receptor for lymphocytic choriomeningitis virus and Lassa fever virus. Science. 1998 Dec 11;282(5396):2079-81. PubMed PMID: 9851928.

330. Raphael AR, Couthouis J, Sakamuri S, Siskind C, Vogel H, Day JW, et al. Congenital muscular dystrophy and generalized epilepsy caused by GMPPB mutations. Brain Res. 2014 Apr 26. PubMed PMID: 24780531.

331. Sharma V, Ichikawa M, He P, Scott DA, Bravo Y, Dahl R, et al. Phosphomannose isomerase inhibitors improve N-glycosylation in selected phosphomannomutase-deficient fibroblasts. J Biol Chem. 2011 Nov 11;286(45):39431-8. PubMed PMID: 21949237. Pubmed Central PMCID: 3234766. Epub 2011/09/29. eng.

332. Koehler K, Malik M, Mahmood S, Giesselmann S, Beetz C, Hennings JC, et al. Mutations in GMPPA cause a glycosylation disorder characterized by intellectual disability and autonomic dysfunction. Am J Hum Genet. 2013 Oct 3;93(4):727-34. PubMed PMID: 24035193. Pubmed Central PMCID: 3791256.

333. Davis AJ, Perugini MA, Smith BJ, Stewart JD, Ilg T, Hodder AN, et al. Properties of GDP-mannose pyrophosphorylase, a critical enzyme and drug target in Leishmania mexicana. J Biol Chem. 2004 Mar 26;279(13):12462-8. PubMed PMID: 14718535. Epub 2004/01/14. eng.

334. Warit S, Zhang N, Short A, Walmsley RM, Oliver SG, Stateva LI. Glycosylation deficiency phenotypes resulting from depletion of GDP-mannose pyrophosphorylase in two yeast species. Mol Microbiol. 2000 Jun;36(5):1156-66. PubMed PMID: 10844699. Epub 2000/06/09. eng.

335. Zhang N, Gardner DC, Oliver SG, Stateva LI. Down-regulation of the expression of PKC1 and SRB1/PSA1/VIG9, two genes involved in cell wall integrity in Saccharomyces cerevisiae, causes flocculation. Microbiology. 1999 Feb;145 (Pt 2):309-16. PubMed PMID: 10075413. Epub 1999/03/13. eng.

336. Jiang H, Ouyang H, Zhou H, Jin C. GDP-mannose pyrophosphorylase is essential for cell wall integrity, morphogenesis and viability of Aspergillus fumigatus. Microbiology. 2008 Sep;154(Pt 9):2730-9. PubMed PMID: 18757806. Epub 2008/09/02. eng.

337. Qin C, Qian W, Wang W, Wu Y, Yu C, Jiang X, et al. GDP-mannose pyrophosphorylase is a genetic determinant of ammonium sensitivity in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2008 Nov 25;105(47):18308-13. PubMed PMID: 19011088. Pubmed Central PMCID: 2587558. Epub 2008/11/18. eng.

338. Keller R, Renz FS, Kossmann J. Antisense inhibition of the GDP-mannose pyrophosphorylase reduces the ascorbate content in transgenic plants leading to developmental changes during senescence. Plant J. 1999 Jul;19(2):131-41. PubMed PMID: 10476060. Epub 1999/09/04. eng.

339. Denton H, Fyffe S, Smith TK. GDP-mannose pyrophosphorylase is essential in the bloodstream form of Trypanosoma brucei. Biochem J. 2010 Feb 1;425(3):603-14. PubMed PMID: 19919534. Epub 2009/11/19. eng.

340. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol. 2014 Mar;32(3):279-84. PubMed PMID: 24463574. Pubmed Central PMCID: 3988262.

341. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007 Mar 23;315(5819):1709-12. PubMed PMID: 17379808.

342. Barrangou R. RNA-mediated programmable DNA cleavage. Nat Biotechnol. 2012 Sep;30(9):836-8. PubMed PMID: 22965054.

343. Faridmoayer A, Fentabil MA, Haurat MF, Yi W, Woodward R, Wang PG, et al. Extreme substrate promiscuity of the Neisseria oligosaccharyl transferase involved in protein O-glycosylation. J Biol Chem. 2008 Dec 12;283(50):34596-604. PubMed PMID: 18930921. Pubmed Central PMCID: 3259870.

344. Gantt RW, Goff RD, Williams GJ, Thorson JS. Probing the aglycon promiscuity of an engineered glycosyltransferase. Angew Chem Int Ed Engl. 2008;47(46):8889-92. PubMed PMID: 18924204. Pubmed Central PMCID: 2963038.

345. Gantt RW, Peltier-Pain P, Singh S, Zhou M, Thorson JS. Broadening the scope of glycosyltransferase-catalyzed sugar nucleotide synthesis. Proc Natl Acad Sci U S A. 2013 May 7;110(19):7648-53. PubMed PMID: 23610417. Pubmed Central PMCID: 3651490.

346. Vuillaumier-Barrot S, Le Bizec C, de Lonlay P, Barnier A, Mitchell G, Pelletier V, et al. Protein losing enteropathy-hepatic fibrosis syndrome in Saguenay-Lac St-Jean, Quebec is a congenital disorder of glycosylation type Ib. J Med Genet. 2002 Nov;39(11):849-51. PubMed PMID: 12414827. Pubmed Central PMCID: 1735008.

347. Niehues R, Hasilik M, Alton G, Korner C, Schiebe-Sukumar M, Koch HG, et al. Carbohydrate-deficient glycoprotein syndrome type Ib. Phosphomannose isomerase

deficiency and mannose therapy. J Clin Invest. 1998 Apr 1;101(7):1414-20. PubMed PMID: 9525984. Pubmed Central PMCID: 508719.

348. Chu J, Mir A, Gao N, Rosa S, Monson C, Sharma V, et al. A zebrafish model of congenital disorders of glycosylation with phosphomannose isomerase deficiency reveals an early opportunity for corrective mannose supplementation. Dis Model Mech. 2012 Sep 6. PubMed PMID: 22899857. Epub 2012/08/18. Eng.

7 Appendices

ID	Gene	Variant type	Forward primer sequence 5'-3'	Reverse primer sequence 5'-3'
F2	*GRIN2A*	*De novo*	GCCAACATACCCAGTAGGC	TTGAGGTCAACGGCATCG
F3 + F16	*PPFIBP2*	*De novo*	CAGGAGTCCAGCCCAGAG	AGACAGCGAGGGCTGTAG
F6	*C11orf41*	*De novo*	ACCTGTGTGCTGATCGAC	AGCAGCTGTTGTGGTAGC
F6	*SMARCC2*	*De novo*	ATTCCCGTGACCCTTTGC	GCCCAGGCTGAAGATGTAG
F6	*NF1*	*De novo*	TGCTCTGTGCAAATGCTTG	AGTCTGCATGGAGTCTGC
F6	*ZHX3*	*De novo*	ACCCCATCTTGCTTGCTG	AAACCACTGCTCCTGCTG
F7	*UNC80*	*De novo*	TGGTTGGAAACCCTTGCC	AGGGCTAGGTGAGAACTCC
F7	*WFDC8*	*De novo*	TTCTGTGGCCCCAAATCC	TGGCACAAGGTACAGCAC
F8	*CD244*	*De novo*	GTAGGCGGGGTTTCTCAAC	AGGTGCTCCTAGGTTCTGG
F9	*PARD3B*	*De novo*	AGGGGAAGCAGCCATTTC	ACGGATCTTCAGTCAGGTG
F10	*ATP6V1B2*	*De novo*	GGGCTACCACACAATGAGG	GGCCAAAATGCCAGATTGC
F10	*SEMA4D*	*De novo*	AGAAGCTCCCTGGCTCTAC	TTCCAAGTGGTCGCCAAG
F14	*STX12*	*De novo*	TAGGAGAGCGGTGTAGAGC	AAACTCGGCTGACCACTG
F14	*KCTD8*	*De novo*	TTCTGCTTGGTGACCTTGG	CTCCGGGAACGAGAAAGTG
F15	*DOCK1*	*De novo*	TAGTGTTGCGGGCTTTCC	GGAAAGGTTGGTCCAGGTC
F18	*FAM3D*	*De novo*	AAGGCAGCACAGCTTGTC	TCTGGAGGGCTAAGTGGAG
F18	*ABCB9*	*De novo*	ACGAACAGGGCCCAAAAC	CCCTTGTGCTGATGTTTGC
F19	*PARD3B*	*De novo*	TGCCCTTCTCCAAACCAC	TCCTGGATCGTTCAAAGGG
F19	*DNAJC13*	*De novo*	TCATGGCCATCACACACG	TGCACATGGCACTGGTTAG
F19	*NLRP1*	*De novo*	CCTCTCCAGAAGCAAGCAG	TCAGGGGAGGAACCTGATG
F20	*COL2A1*	*De novo*	TGTTAGCTGCAGGCTGATG	CTGGCTCATGTGCCTATGG
F22	*TACR2*	*De novo*	CATGGCTGTGATGGGGAAG	ACTGGGCTTTGTGCTCAG
F23	*FGFR3*	*De novo*	ACTGGCGTTACTGACTGC	TTCGTGCCCCAAAGTACC
F25	*SMARCC1*	*De novo*	GCAGAAAGGCACAACCTG	GGACAAGGAAGAAGCAAGC
F25	*PNLIPRP1*	*De novo*	GAGAGAGAGGCAGAGAAGC	CCCTTGCCAGAAATGTGC
F26	*KDM5B*	*De novo*	GCAAAGGAAGGCTGTTTAGC	GTCTTACAAAGCCGAGTCTG
F26	*STAU2*	*De novo*	AAACCGAATGCAGCCGAG	AGTCAGTGAATGGCTCTCTC
F27	*C2orf40*	*De novo*	TATTCTTCGCCCCAACTCC	GCCCAGTTTTGTAGCTTGC
F27	*INSC*	*De novo*	AAGGCATGGAGGAACAGC	TCGGCCCCAAGTTACAAAC
F28	*PPP6R1*	*De novo*	GTGGAAGCTTGGAGAAGGG	TGGTTCGGGTTGTGTGTG
F31	*FMNL3*	*De novo*	ACCAATGCCACCTCATGC	AGCTAACTCCCCAGTCAGG
F33	*SEC31B*	*De novo*	AGGGAGTGGGTAGGGAAAG	TTCCTGGTTCCCCTCTACC
F33	*EGFL6*	*De novo*	TGGCAGGTCACAAGAAAGAC	TTTGGAAGGACGCTGGTG
F1	*PRKDC*	Het, M	TAGGAGTTCAAAAGTTGTGTCAA	GCTTGGGATAGAATTGCACC

F1	*PRKDC*	Het, P	GAGGCTTTCTGGAGAGTTTTG	TGCTATCCAGCAGCTTCC
F5	*DLC1*	Het, M	ACTGCCATTGGTGAGAGC	GCAACTTGGCAGGCAATG
F5	*DLC1*	Het, P	TGCCATCTTCTGCCTTGAC	AGAAAAGGCACTGCCCATC
F6	*RERE*	Het, M	AAAGTGCTCCATGGGGTTG	AGTGTGGACCAAGCGGA
F6	*RERE*	Het, P	AAGGCCACATGACACTGC	AGCTACCCTGACAGACTGG
F10	*MACF1*	Het, M	AGGAGTGGTTGGATTGCC	CCACCATTTCCCCTTCCTC
F10	*MACF1*	Het, P	GTCCACCAGCCAAGTACAG	CTCACTGCAACTAGGAGAAGG
F13	*FRAS1*	Het, M	TGGGTAAACGGCCATGTG	TGCTCAGCAGTGTCATTACC
F13	*FRAS1*	Het, P	TCCCTGTCAGAAGACCGAG	CTCTCAGGGGCTGTGAAAC
F8	*RNF213*	Het, M	CTCACCTGGTGTAGTGCAG	AATCCCAACGTGGGTGTG
F8	*RNF213*	Het, P	TTCGGCGACTTCGTCTC	CTCACCTGGTGTAGTGCAG
F12	*DACH1*	Hom, M+P	TCAGGAACAGGTCGAAAGC	TTCGGCGACTTCGTCTC

Appendix 1: Primer sequences for Sanger sequencing of variants that passed validation. Het = heterozygous; Hom = homozygous; M = maternally inherited; P = paternally inherited.

ID	SEX	GT	CHR	POS	REF	ALT	Gene	CQ
F1	M	Hom	4	1389005	T	C	*CRIPAK*	NS
F1	M	Comp het	8	48719844	G	A	*PRKDC*	NS
F1	M	Comp het	8	48848319	C	A	*PRKDC*	NS
F1	M	Comp het	8	101718965	G	A	*PABPC1*	NS
F1	M	Comp het	8	101718968	C	T	*PABPC1*	NS
F1	M	Comp het	8	101719138	C	T	*PABPC1*	NS
F1	M	Comp het	8	101719201	A	G	*PABPC1*	NS
F1	M	Comp het	8	10466003	GCTGGGCCTCCCCTTCAGCCTC	G	*RP1L1*	NS
F1	M	Comp het	8	10469454	G	T	*RP1L1*	NS
F1	M	Comp het	11	93808384	A	G	*HEPHL1*	NS
F1	M	Comp het	11	93806297	C	G	*HEPHL1*	NS
F1	M	Comp het	13	100622837	C	CCGG	*ZIC5*	NS
F1	M	Comp het	13	100622667	TGGC	T	*ZIC5*	NS
F1	M	Hom	14	24769849	A	AGAGGAG	*C14orf21*	NS
F1	M	Comp het	19	12358606	G	A	*ZNF44*	NS
F1	M	Comp het	19	12383893	G	A	*ZNF44*	NS
F1	M	Hemi	X	129146962	C	T	*BCORL1*	NS
F1	M	Hemi	X	34149726	G	A	*FAM47A*	NS
F1	M	Hemi	X	108868195	C	A	*KCNE1L*	STOP
F1	M	Hemi	X	117960384	G	A	*ZCCHC12*	NS
F2	F	Comp het	1	222802423	G	A	*MIA3*	NS
F2	F	Comp het	1	222802652	T	C	*MIA3*	NS
F2	F	Comp het	4	37445867	C	T	*KIAA1239*	NS
F2	F	Comp het	4	37446545	C	A	*KIAA1239*	NS
F2	F	Comp het	6	29912028	AG	A	*HLA-A*	FS
F2	F	Comp het	6	29911063	T	G	*HLA-A*	NS
F2	F	Comp het	9	90502542	T	C	*C9orf79*	NS
F2	F	Comp het	9	90500202	A	G	*C9orf79*	NS
F2	F	Hom	11	18127558	C	CCGG	*SAAL1*	NS
F2	F	Comp het	16	1470583	C	G	*C16orf91*	NS
F2	F	Comp het	16	1476330	T	C	*C16orf91*	NS
F2	F	Comp het	17	20799179	C	G	*CCDC144NL*	NS
F2	F	Comp het	17	20799281	G	A	*CCDC144NL*	NS
F2	F	Comp het	19	49113161	C	T	*FAM83E*	NS
F2	F	Comp het	19	49113215	G	A	*FAM83E*	NS
F2	F	Comp het	20	36870301	C	T	*KIAA1755*	NS
F2	F	Comp het	20	36868106	G	A	*KIAA1755*	NS
F2	F	Hom	21	34003928	A	AAGTATT	*SYNJ1*	NS
F3	M	Comp het	7	142561747	C	T	*EPHB6*	STOP
F3	M	Comp het	7	142562051	C	CCCTCCT	*EPHB6*	NS
F3	M	Comp het	1	17084536	TGGAACA	T	*MST1P9*	NS
F3	M	Comp het	1	17085427	T	C	*MST1P9*	NS
F3	M	Comp het	1	17087582	G	A	*MST1P9*	NS
F3	M	Comp het	6	138752868	C	A	*NHSL1*	NS
F3	M	Comp het	6	138794490	G	A	*NHSL1*	NS
F3	M	Hom	14	74060513	GCTTA	G	*ACOT4*	FS
F3	M	Hom	12	7045891	A	ACAG	*ATN1*	NS
F3	M	Hemi	X	49104709	C	T	*CCDC22*	NS
F3	M	Hom	19	54973988	GCCT	G	*LENG9*	NS
F3	M	Hom	12	124887058	G	GGCTGCT, GGCT	*NCOR2*	NS
F3	M	Hemi	X	9863050	C	T	*SHROOM2*	NS
F3	M	Hom	9	139277994	CGCT	C	*SNAPC4*	NS
F3	M	Hom	2	231861032	TCAGCAGCCTAGCCCTGAATCCACACCA	T	*SPATA3*	INDEL
F5	M	Comp het	8	12957657	C	T	*DLC1*	NS
F5	M	Comp het	8	13356860	G	C	*DLC1*	NS
F5	M	Comp het	9	139276718	C	T	*SNAPC4*	NS
F5	M	Comp het	9	139277994	CGCT	C	*SNAPC4*	NS
F5	M	Comp het	2	179430460	A	G	*TTN*	NS
F5	M	Comp het	2	179497758	A	G	*TTN*	NS
F5	M	Comp het	2	179579172	C	T	*TTN*	NS
F5	M	Comp het	2	179634421	T	G	*TTN*	NS

F5	M	Hom	21	28215826	C	CACA	*ADAMTS1*	NS
F5	M	Hemi	X	119394834	G	A	*FAM70A*	NS
F5	M	Hemi	X	31089928	G	A	*FTHL17*	NS
F5	M	Hemi	X	135430934	C	A	*GPR112*	NS
F5	M	Hemi	X	108652306	C	T	*GUCY2F*	NS
F5	M	Hom	6	33052736	G	A	*HLA-DPB1*	NS
F5	M	Hom	2	170632960	C	CA,CAAA,CAA	*KLHL23*	FS
F5	M	Hemi	X	99551442	G	C	*PCDH19*	NS
F5	M	Hemi	X	114426292	G	A	*RBMXL3*	NS
F5	M	Hom	18	42456670	C	CTCTT	*SETBP1*	FS
F5	M	Hom	21	34003928	A	AAGTATT	*SYNJ1*	NS
F5	M	Hemi	X	117528073	A	C	*WDR44*	NS
F6	F	Comp het	10	134663933	A	G	*C10orf93*	NS
F6	F	Comp het	10	134680995	C	T	*C10orf93*	NS
F6	F	Comp het	10	134726261	A	C	*C10orf93*	NS
F6	F	Comp het	10	134686163	A	G	*C10orf93*	NS
F6	F	Comp het	10	134694526	A	G	*C10orf93*	NS
F6	F	Comp het	7	30818142	T	G	*FAM188B*	NS
F6	F	Comp het	7	30825544	C	A	*FAM188B*	NS
F6	F	Comp het	7	103205827	G	C	*RELN*	NS
F6	F	Comp het	7	103141235	G	A	*RELN*	NS
F6	F	Comp het	1	8418331	C	T	*RERE*	NS
F6	F	Comp het	1	8418909	C	T	*RERE*	NS
F6	F	Hom	9	95237024	CTCA	C	*ASPN*	NS
F6	F	Hom	19	41754430	G	A	*AXL*	NS
F6	F	Hom	7	100550191	C	T	*MUC3A*	NS
F6	F	Hom	6	1611802	G	GGGC	*FOXC1*	NS
F6	F	Hom	14	106329450	T	TACC	*IGHJ6*	NS
F6	F	Hom	6	32191658	TAGCAGCAGCAGCAGC	TAGCAGCAGCAGCAGCAGCAGC,T	*NOTCH4*	NS
F6	F	Hom	6	41754415	G	A	*PRICKLE4*	NS
F6	F	Hom	19	43702335	C	G	*PSG4*	NS
F6	F	Hom	3	52027853	T	TCCTTGG	*RPL29*	NS
F6	F	Hom	22	39777822	C	CCAA	*SYNGR1*	NS
F6	F	Hom	18	3452222	CT	C	*TGIF1*	FS
F7	F	Comp het	16	55862791	T	C	*CES1*	NS
F7	F	Comp het	16	55862824	C	T	*CES1*	NS
F7	F	Comp het	4	1388757	T	C	*CRIPAK*	NS
F7	F	Comp het	4	1389005	T	C	*CRIPAK*	NS
F7	F	Comp het	19	9018166	A	G	*MUC16*	NS
F7	F	Comp het	19	9082960	G	A	*MUC16*	NS
F7	F	Comp het	13	45149973	G	A	*TSC22D1*	NS
F7	F	Comp het	13	45148705	TTGC	T	*TSC22D1*	NS
F7	F	Comp het	2	179399071	G	A	*TTN*	NS
F7	F	Comp het	2	179641112	C	A	*TTN*	NS
F7	F	Comp het	2	179431633	C	T	*TTN*	NS
F7	F	Hom	15	74536400	TAAG	T	*CCDC33*	NS
F7	F	Hom	8	8234868	C	CGCCGCT	*SGK223*	NS
F7	F	Hom	6	32549611	T	A	*HLA-DRB1*	NS
F7	F	Hom	19	43708978	TC	T	*PSG4*	FS
F7	F	Hom	3	52027853	T	TCCTTGG	*RPL29*	NS
F8	M	Comp het	2	160738803	G	A	*LY75-CD302*	NS
F8	M	Comp het	2	160688217	T	C	*LY75-CD302*	NS
F8	M	Comp het	3	195452754	CAGAAAT	C	*MUC20*	NS
F8	M	Comp het	3	195346656	G	A	*MUC20*	NS
F8	M	Comp het	3	195447886	G	C	*MUC20*	NS
F8	M	Comp het	17	78264463	AGAG	A	*RNF213*	NS
F8	M	Comp het	17	78357600	A	G	*RNF213*	NS
F8	M	Comp het	2	179399677	C	T	*TTN*	NS
F8	M	Comp het	2	179412829	C	T	*TTN*	NS
F8	M	Comp het	2	179591953	C	G	*TTN*	NS
F8	M	Comp het	16	74937918	C	T	*WDR59*	NS
F8	M	Comp het	16	74990380	G	A	*WDR59*	NS
F8	M	Hom	17	48452978	A	AAGC	*EME1*	NS
F8	M	Hom	19	40392585	T	G	*FCGBP*	NS
F8	M	Hom	14	23744800	ACAT	A	*HOMEZ*	NS

F8	M	Hom	19	55790886	A	AGCCGCCGCC	*HSPBP1*	NS
F8	M	Hom	16	71956511	AATGCCC	A	*KIAA0174*	NS
F8	M	Hom	7	100635591	C	A	*MUC12*	NS
F8	M	Hemi	X	153040414	C	T	*PLXNB3*	STOP
F8	M	Hemi	X	16870553	C	T	*RBBP7*	NS
F8	M	Hemi	X	50350728	T	TTCC	*SHROOM4*	NS
F8	M	Hemi	X	99920314	G	T	*SRPX2*	NS
F8	M	Hom	22	44258311	ACGCGCC	A	*SULT4A1*	INDEL
F8	M	Hom	19	44589999	TCTC	T	*ZNF284*	NS
F9	M	Comp het	7	48318614	T	G	*ABCA13*	NS
F9	M	Comp het	7	48547481	C	T	*ABCA13*	NS
F9	M	Comp het	2	202352480	G	A	*ALS2CR11*	NS
F9	M	Comp het	2	202400832	C	T	*ALS2CR11*	NS
F9	M	Comp het	3	130300740	C	T	*COL6A6*	STOP
F9	M	Comp het	3	130381038	G	A	*COL6A6*	NS
F9	M	Comp het	8	144942321	C	T	*EPPK1*	NS
F9	M	Comp het	8	144940230	G	C	*EPPK1*	NS
F9	M	Comp het	20	57430118	C	G	*GNAS*	NS
F9	M	Comp het	20	57428948	G	C	*GNAS*	NS
F9	M	Comp het	7	100679024	A	G	*MUC17*	NS
F9	M	Comp het	7	100685477	A	T	*MUC17*	NS
F9	M	Comp het	11	1080917	C	G	*MUC2*	NS
F9	M	Comp het	11	1093600	G	A	*MUC2*	NS
F9	M	Comp het	20	61288233	G	A	*SLCO4A1*	NS
F9	M	Comp het	20	61299350	G	A	*SLCO4A1*	NS
F9	M	Comp het	16	2806466	C	T	*SRRM2*	NS
F9	M	Comp het	16	2817604	G	A	*SRRM2*	NS
F9	M	Comp het	16	2817749	C	G	*SRRM2*	NS
F9	M	Comp het	2	234878910	C	T	*TRPM8*	NS
F9	M	Comp het	2	234891850	G	A	*TRPM8*	NS
F9	M	Hom	11	130298117	GGCA	G	*ADAMTS8*	NS
F9	M	Hemi	X	152814163	G	A	*ATP2B3*	NS
F9	M	Hemi	X	49103316	G	A	*CCDC22*	NS
F9	M	Hom	13	46170719	CCCAGATACTCTTCCTCCT	C	*FAM194B*	NS
F9	M	Hom	19	55790886	A	AGCCGCCGCC,AGCCGCC	*HSPBP1*	NS
F9	M	Hom	4	4276475	C	T	*LYAR*	NS
F9	M	Hom	19	56029616	C	CCCA	*SSC5D*	NS
F10	F	Comp het	10	85961593	C	T	*CDHR1*	NS
F10	F	Comp het	10	85955337	C	A	*CDHR1*	NS
F10	F	Comp het	12	124413109	T	C	*DNAH10*	NS
F10	F	Comp het	12	124330648	G	A	*DNAH10*	NS
F10	F	Comp het	6	139222224	C	T	*ECT2L*	NS
F10	F	Comp het	6	139202137	T	C	*ECT2L*	NS
F10	F	Comp het	1	39851427	G	A	*MACF1*	NS
F10	F	Comp het	1	39901245	A	G	*MACF1*	NS
F10	F	Comp het	8	10467605	C	T	*RP1L1*	NS
F10	F	Comp het	8	10467652	G	C	*RP1L1*	NS
F10	F	Hom	12	8374781	C	CACG	*FAM90A1*	NS
F10	F	Hom	9	112900341	G	GGAAGCT	*PALM2-AKAP2*	NS
F11	M	Comp het	6	32552059	G	T	*HLA-DRB1*	NS
F11	M	Comp het	6	32557506	T	C	*HLA-DRB1*	NS
F11	M	Comp het	4	57777171	C	G	*REST*	NS
F11	M	Comp het	4	57796913	C	T	*REST*	NS
F11	M	Hemi	X	71521598	G	A	*CITED1*	NS
F11	M	Hom	4	3590823	GACACAC	GACAC,G	*RP3-368B9.1*	FS
F11	M	Hom	8	95272605	G	C	*GEM*	NS
F11	M	Hemi	X	3242339	T	C	*MXRA5*	NS
F11	M	Hom	1	1684347	C	CCCT	*NADK*	NS
F11	M	Hom	12	124887058	G	GGCTGCT	*NCOR2*	NS
F11	M	Hemi	X	30322699	T	C	*NR0B1*	NS
F12	F	Comp het	7	48626765	A	G	*ABCA13*	NS
F12	F	Comp het	7	48314151	C	G	*ABCA13*	NS
F12	F	Comp het	21	47852049	G	A	*PCNT*	NS
F12	F	Comp het	21	47831695	G	A	*PCNT*	NS

F12	F	Comp het	2	179414177	G	A	*TTN*	NS
F12	F	Comp het	2	179486037	C	A	*TTN*	NS
F12	F	Comp het	2	179396782	C	G	*TTN*	NS
F12	F	Comp het	2	179484593	C	T	*TTN*	NS
F12	F	Comp het	2	179599473	C	G	*TTN*	NS
F12	F	Comp het	6	56999585	C	A	*ZNF451*	NS
F12	F	Comp het	6	57012673	C	T	*ZNF451*	NS
F12	F	Hom	14	74060511	T	TTCAA	*ACOT4*	FS
F12	F	Hom	1	111833488	C	CCT	*CHIA*	FS
F12	F	Hom	9	39171471	C	T	*CNTNAP3*	NS
F12	F	Hom	13	72440658	TGCCGCC	T	*DACH1*	NS
F12	F	Hom	6	32191658	TAGC	T	*NOTCH4*	NS
F12	F	Hom	19	43708978	TC	T	*PSG4*	FS
F12	F	Hom	7	99662511	GTAGT	G	*ZNF3*	FS
F13	M	Comp het	4	79238620	C	T	*FRAS1*	NS
F13	M	Comp het	4	79353746	C	A	*FRAS1*	NS
F13	M	Comp het	15	42145586	G	A	*SPTBN5*	NS
F13	M	Comp het	15	42154034	C	T	*SPTBN5*	NS
F13	M	Comp het	2	1544464	C	T	*TPO*	NS
F13	M	Comp het	2	1459885	A	G	*TPO*	NS
F13	M	Hemi	X	110980029	G	C	*ALG13*	NS
F13	M	Hemi	X	134713929	C	G	*DDX26B*	NS
F13	M	Hom	19	46815703	ATATT	A	*HIF3A*	FS
F13	M	Hom	19	49657710	ACAT	A	*HRC*	NS
F13	M	Hom	14	106329450	T	TACC	*IGHJ6*	NS
F13	M	Hemi	X	135314244	G	A	*MAP7D3*	NS
F13	M	Hom	2	178494173	G	GGGA	*PDE11A*	NS
F13	M	Hom	16	81242148	GTT	G	*PKD1L2*	FS
F13	M	Hom	9	98270320	AGTGAGTGT	A	*PTCH1*	INDEL
F13	M	Hom	18	3452222	CT	C	*TGIF1*	FS
F13	M	Hemi	X	12904292	T	A	*TLR7*	NS
F13	M	Hom	3	42251577	C	CGGA	*TRAK1*	NS
F14	F	Comp het	20	49508015	T	C	*ADNP*	NS
F14	F	Comp het	20	49508508	C	T	*ADNP*	NS
F14	F	Comp het	2	242162665	C	T	*ANO7*	NS
F14	F	Comp het	2	242144345	T	G	*ANO7*	NS
F14	F	Comp het	1	214819026	A	C	*CENPF*	NS
F14	F	Comp het	1	214818291	G	A	*CENPF*	NS
F14	F	Comp het	4	155156575	T	C	*DCHS2*	NS
F14	F	Comp het	4	155219318	C	T	*DCHS2*	NS
F14	F	Comp het	1	17087541	GGTGCT	G	*MST1P9*	FS
F14	F	Comp het	1	17085427	T	C	*MST1P9*	NS
F14	F	Comp het	6	46661479	G	T	*TDRD6*	NS
F14	F	Comp het	6	46660511	T	A	*TDRD6*	NS
F14	F	Hom	6	109850199	AAC	A	*AKD1*	FS
F14	F	Hom	12	8374781	C	CACG	*FAM90A1*	NS
F14	F	Hom	14	106878056	T	G	*IGHV4-39*	NS
F14	F	Hom	9	78790153	A	G	*PCSK5*	NS
F14	F	Hom	1	12919963	T	G	*PRAMEF2*	NS
F15	F	Comp het	5	148627397	C	T	*ABLIM3*	NS
F15	F	Comp het	5	148586585	A	G	*ABLIM3*	NS
F15	F	Comp het	5	82833426	A	G	*VCAN*	NS
F15	F	Comp het	5	82835589	T	C	*VCAN*	NS
F15	F	Hom	6	157100396	G	GCGC	*ARID1B*	NS
F15	F	Hom	1	111833488	C	CCT	*CHIA*	FS
F15	F	Hom	2	241696840	ATCC	A	*KIF1A*	NS
F15	F	Hom	19	54973988	GCCT	G	*LENG9*	NS
F15	F	Hom	1	1684347	C	CCCT	*NADK*	NS
F16	M	Comp het	16	1470583	C	G	*C16orf91*	NS
F16	M	Comp het	16	1476330	T	C	*C16orf91*	NS
F16	M	Comp het	9	90502542	T	C	*C9orf79*	NS
F16	M	Comp het	9	90500202	A	G	*C9orf79*	NS
F16	M	Comp het	17	20799179	C	G	*CCDC144NL*	NS
F16	M	Comp het	17	20799281	G	A	*CCDC144NL*	NS
F16	M	Comp het	7	142561747	C	T	*EPHB6*	STOP
F16	M	Comp het	7	142562051	C	CCCTCCT	*EPHB6*	NS
F16	M	Comp het	1	17084536	TGGAACA	T	*MST1P9*	NS
F16	M	Comp het	1	17085427	T	C	*MST1P9*	NS
F16	M	Comp het	6	138752868	C	A	*NHSL1*	NS

F16	M	Comp het	6	138794490	G	A	*NHSL1*	NS
F16	M	Hom	9	95237024	CTCA	CTCATCA,C	*ASPN*	NS
F16	M	Hom	12	7045891	A	ACAG	*ATN1*	NS
F16	M	Hom	10	46999591	C	CATGAGGGAG	*GPRIN2*	NS
F16	M	Hom	19	54973988	GCCT	G	*LENG9*	NS
F16	M	Hom	12	124887058	G	GGCT	*NCOR2*	NS
F16	M	Hemi	X	9863050	C	T	*SHROOM2*	NS
F16	M	Hemi	X	50350728	T	TTCC	*SHROOM4*	NS
F16	M	Hom	9	139277994	CGCT	C	*SNAPC4*	NS
F16	M	Hom	2	231861032	TCAGCAGCCTAGCCCTGAATCCACACCA	T	*SPATA3*	INDEL
F17	F	Comp het	16	2369688	A	T	*ABCA3*	NS
F17	F	Comp het	16	2374481	T	C	*ABCA3*	NS
F17	F	Comp het	13	42876835	T	G	*AKAP11*	NS
F17	F	Comp het	13	42875878	C	T	*AKAP11*	NS
F17	F	Comp het	1	68960131	T	C	*DEPDC1*	NS
F17	F	Comp het	1	68960186	T	C	*DEPDC1*	NS
F17	F	Comp het	16	71950412	G	A	*KIAA0174*	NS
F17	F	Comp het	16	71956511	AATGCCC	A	*KIAA0174*	NS
F17	F	Comp het	1	26303228	G	A	*PAFAH2*	NS
F17	F	Comp het	1	26317303	C	T	*PAFAH2*	NS
F17	F	Comp het	7	75070377	T	A	*POM121C*	NS
F17	F	Comp het	7	75070840	C	A	*POM121C*	NS
F17	F	Hom	20	11830180	CT	C	*C20orf61*	FS
F17	F	Hom	13	46170719	CCCAGATACTCTTCCTCCT	C	*FAM194B*	NS
F17	F	Hom	7	15725797	ATGG	A	*MEOX2*	NS
F17	F	Hom	9	98270320	AGTGAGTGT	A	*PTCH1*	INDEL
F17	F	Hom	18	42456670	C	CTCTT	*SETBP1*	FS
F17	F	Hom	22	50468907	C	G	*TTLL8*	NS
F18	M	Comp het	3	136019898	C	T	*PCCB*	NS
F18	M	Comp het	3	135969390	A	C	*PCCB*	NS
F18	M	Comp het	2	179452447	T	C	*TTN*	NS
F18	M	Comp het	2	179611552	C	T	*TTN*	NS
F18	M	Comp het	16	72831629	G	A	*ZFHX3*	NS
F18	M	Comp het	16	72831357	C	CTTGTTG	*ZFHX3*	NS
F18	M	Comp het	16	72832550	A	C	*ZFHX3*	NS
F18	M	Hemi	X	105855323	T	C	*CXorf57*	NS
F18	M	Hemi	X	44703940	A	G	*DUSP21*	NS
F18	M	Hemi	X	138644189	C	T	*F9*	NS
F18	M	Hemi	X	55650995	C	T	*FOXR2*	NS
F18	M	Hemi	X	131842557	G	C	*HS6ST2*	NS
F18	M	Hom	16	71956511	AATGCCC	A	*KIAA0174*	NS
F18	M	Hom	1	201356001	CCCA	C	*LAD1*	NS
F18	M	Hom	5	112824048	T	TGCC	*MCC*	NS
F18	M	Hemi	X	119077233	C	G	*NKAP*	NS
F18	M	Hemi	X	129546514	G	A	*RBMX2*	NS
F18	M	Hemi	X	50350728	T	TTCC	*SHROOM4*	NS
F18	M	Hom	18	3452222	CT	C	*TGIF1*	FS
F19	M	Comp het	14	105415079	G	T	*AHNAK2*	NS
F19	M	Comp het	14	105419557	G	A	*AHNAK2*	NS
F19	M	Comp het	12	7527284	C	T	*CD163L1*	NS
F19	M	Comp het	12	7521535	A	G	*CD163L1*	NS
F19	M	Comp het	3	52409413	C	G	*DNAH1*	NS
F19	M	Comp het	3	52426643	G	A	*DNAH1*	NS
F19	M	Comp het	5	13883075	C	T	*DNAH5*	NS
F19	M	Comp het	5	13759007	G	A	*DNAH5*	NS
F19	M	Comp het	2	84880481	C	G	*DNAH6*	NS
F19	M	Comp het	2	84924743	C	G	*DNAH6*	NS
F19	M	Comp het	5	132534965	G	A	*FSTL4*	NS
F19	M	Comp het	5	132939589	C	T	*FSTL4*	NS
F19	M	Comp het	16	87637893	C	CCTGCTG	*JPH3*	NS
F19	M	Comp het	16	87723683	G	A	*JPH3*	NS
F19	M	Comp het	16	71712805	C	A	*PHLPP2*	NS
F19	M	Comp het	16	71724598	T	C	*PHLPP2*	NS

F19	M	Hom	16	84229207	C	T	*ADAD2*	NS
F19	M	Hom	12	121093629	CGTGCGT	C	*CABP1*	NS
F19	M	Hemi	X	107431191	T	A	*COL4A6*	NS
F19	M	Hemi	X	2793951	T	C	*GYG2*	NS
F19	M	Hom	17	39254335	A	AT	*KRTAP4-8*	FS
F19	M	Hemi	X	140993827	A	C	*MAGEC1*	NS
F19	M	Hemi	X	140994031	G	T	*MAGEC1*	NS
F19	M	Hom	12	124887058	G	GGCT	*NCOR2*	NS
F19	M	Hom	21	47831802	C	T	*PCNT*	NS
F19	M	Hemi	X	152225801	G	A	*PNMA3*	NS
F19	M	Hemi	X	84362764	G	A	*SATL1*	NS
F19	M	Hemi	X	9863131	G	A	*SHROOM2*	NS
F19	M	Hom	18	3452222	CT	C	*TGIF1*	FS
F19	M	Hom	3	42251577	CGGA	C	*TRAK1*	NS
F20	M	Comp het	8	61778448	C	T	*CHD7*	NS
F20	M	Comp het	8	61769198	C	G	*CHD7*	NS
F20	M	Comp het	6	131277390	G	A	*EPB41L2*	NS
F20	M	Comp het	6	131247845	T	A	*EPB41L2*	NS
F20	M	Comp het	1	152280900	T	G	*FLG*	NS
F20	M	Comp het	1	152281007	A	G	*FLG*	NS
F20	M	Comp het	5	90002053	A	G	*GPR98*	NS
F20	M	Comp het	5	90059270	C	A	*GPR98*	NS
F20	M	Comp het	6	32729613	T	C	*HLA-DQB2*	NS
F20	M	Comp het	6	32725618	ATG	A	*HLA-DQB2*	FS
F20	M	Comp het	6	32182013	C	T	*NOTCH4*	NS
F20	M	Comp het	6	32191658	T	TAGC	*NOTCH4*	NS
F20	M	Comp het	1	12337667	C	T	*VPS13D*	NS
F20	M	Comp het	1	12378274	C	T	*VPS13D*	NS
F20	M	Comp het	1	12337460	T	A	*VPS13D*	NS
F20	M	Hom	17	42981239	G	GAGT	*FAM187A*	NS
F20	M	Hemi	X	152860096	C	T	*FAM58A*	NS
F20	M	Hom	6	31237727	T	C	*HLA-C*	NS
F20	M	Hom	6	31239622	C	A	*HLA-C*	NS
F20	M	Hom	6	32632781	A	T	*HLA-DQB1*	NS
F20	M	Hom	16	87637893	C	CCTGCTG	*JPH3*	NS
F20	M	Hemi	X	154290176	C	G	*MTCP1NB*	NS
F20	M	Hemi	X	153697736	C	T	*PLXNA3*	NS
F20	M	Hom	21	40883671	G	GAGA	*SH3BGR*	NS
F20	M	Hemi	X	153716622	T	C	*SLC10A3*	NS
F20	M	Hom	11	6411930	CCTGGTGCTGGCG	C	*SMPD1*	NS
F21	M	Comp het	16	1265315	G	A	*CACNA1H*	NS
F21	M	Comp het	16	1270350	G	A	*CACNA1H*	NS
F21	M	Comp het	1	36203016	C	T	*CLSPN*	NS
F21	M	Comp het	1	36212547	G	T	*CLSPN*	NS
F21	M	Comp het	6	33048529	C	T	*HLA-DPB1*	NS
F21	M	Comp het	6	33052736	G	A	*HLA-DPB1*	NS
F21	M	Comp het	6	51656129	C	G	*PKHD1*	NS
F21	M	Comp het	6	51768399	A	T	*PKHD1*	NS
F21	M	Hemi	X	100911707	G	A	*ARMCX2*	NS
F21	M	Hemi	X	65824281	G	A	*EDA2R*	NS
F21	M	Hom	4	3590823	GAC	G	*RP3-368B9.1*	INDEL
F21	M	Hom	7	100550191	C	T	*MUC3A*	NS
F21	M	Hemi	X	135593768	G	A	*HTATSF1*	NS
F21	M	Hom	14	104641986	C	G	*KIF26A*	NS
F21	M	Hemi	X	135303057	T	C	*MAP7D3*	NS
F21	M	Hom	6	31379931	G	A	*MICA*	NS
F21	M	Hemi	X	63490871	TC	T	*MTMR8*	FS
F21	M	Hemi	X	3239828	T	C	*MXRA5*	NS
F21	M	Hom	10	27702256	G	GC	*PTCHD3*	FS
F21	M	Hom	3	42251577	C	CGGA	*TRAK1*	NS
F22	M	Comp het	8	91057198	A	G	*DECR1*	NS
F22	M	Comp het	8	91049129	C	G	*DECR1*	NS
F22	M	Comp het	15	45412435	G	A	*DUOXA1*	NS
F22	M	Comp het	15	45411495	C	A	*DUOXA1*	NS
F22	M	Comp het	3	13679659	T	G	*FBLN2*	NS
F22	M	Comp het	3	13612786	G	A	*FBLN2*	NS
F22	M	Comp het	2	152346522	G	A	*NEB*	NS
F22	M	Comp het	2	152384078	C	T	*NEB*	NS

F22	M	Comp het	4	1066820	C	A	*RNF212*	STOP
F22	M	Comp het	4	1087327	G	GCTGCCCAGGCTGGAGCCAGCC	*RNF212*	NS
F22	M	Comp het	15	62212467	C	T	*VPS13C*	NS
F22	M	Comp het	15	62212770	T	C	*VPS13C*	NS
F22	M	Hom	14	106329450	T	TACC	*IGHJ6*	NS
F22	M	Hom	16	71956511	AATGCCC	A	*KIAA0174*	NS
F22	M	Hemi	X	135314244	G	A	*MAP7D3*	NS
F22	M	Hom	5	140553876	T	C	*PCDHB7*	NS
F22	M	Hom	21	47754510	A	G	*PCNT*	NS
F22	M	Hom	11	209894	ACCC	A	*RIC8A*	NS
F22	M	Hom	2	128744480	T	C	*SAP130*	NS
F22	M	Hom	11	117789312	CGGGCTGGAGATGCCT	C	*TMPRSS13*	INDEL
F22	M	Hemi	X	70466490	G	C	*ZMYM3*	NS
F23	M	Comp het	1	170961328	C	T	*C1orf129*	NS
F23	M	Comp het	1	170952626	T	C	*C1orf129*	NS
F23	M	Comp het	15	22969250	C	T	*CYFIP1*	NS
F23	M	Comp het	15	22925782	G	C	*CYFIP1*	NS
F23	M	Comp het	17	7671351	G	A	*DNAH2*	NS
F23	M	Comp het	17	7663119	C	T	*DNAH2*	NS
F23	M	Comp het	11	70336479	C	T	*SHANK2*	NS
F23	M	Comp het	11	70332311	C	T	*SHANK2*	NS
F23	M	Comp het	2	179404498	G	C	*TTN*	NS
F23	M	Comp het	2	179424272	C	A	*TTN*	NS
F23	M	Comp het	2	179454530	C	T	*TTN*	NS
F23	M	Comp het	2	179610967	C	T	*TTN*	NS
F23	M	Hemi	X	112022297	C	CAGG	*AMOT*	NS
F23	M	Hom	12	8374781	C	CACG	*FAM90A1*	NS
F23	M	Hom	5	74018232	A	G	*GFM2*	NS
F23	M	Hom	1	117122285	G	GTCC	*IGSF3*	NS
F23	M	Hom	15	71276480	GCAA	G	*LRRC49*	NS
F23	M	Hemi	X	19398315	C	T	*MAP3K15*	NS
F23	M	Hemi	X	135313855	T	C	*MAP7D3*	NS
F23	M	Hom	15	100252709	CCAGCAG	C,CCAG	*MEF2A*	NS
F23	M	Hom	2	231861032	TCAGCAGCCTAGCCCTGAATCCACACCA	T	*SPATA3*	INDEL
F23	M	Hemi	X	132161879	G	GTGT	*USP26*	NS
F25	M	Comp het	1	156497776	C	CA	*IQGAP3*	FS
F25	M	Comp het	1	156504308	G	A	*IQGAP3*	NS
F25	M	Comp het	2	179539777	A	G	*TTN*	NS
F25	M	Comp het	2	179634421	T	G	*TTN*	NS
F25	M	Hemi	X	39932564	C	T	*BCOR*	NS
F25	M	Hom	16	89017433	C	T	*RP11-830F9.6*	NS
F25	M	Hom	19	46815703	ATATT	A	*HIF3A*	FS
F25	M	Hom	14	106329450	T	TACC	*IGHJ6*	NS
F25	M	Hemi	X	48823056	G	C	*KCND1*	NS
F25	M	Hom	9	125391770	C	CA,CAA	*OR1B1*	FS
F25	M	Hom	9	78790153	A	G	*PCSK5*	NS
F25	M	Hom	16	81242148	GTT	G	*PKD1L2*	FS
F25	M	Hom	19	11558340	AGAG	A	*PRKCSH*	NS
F25	M	Hom	4	152201018	G	GCAGGT	*PRSS48*	FS
F25	M	Hemi	X	102755132	TC	T	*RAB40A*	FS
F25	M	Hemi	X	132160102	G	A	*USP26*	NS
F26	M	Comp het	1	145515696	A	T	*GNRHR2*	NS
F26	M	Comp het	1	145515394	A	G	*GNRHR2*	NS
F26	M	Comp het	16	72107691	G	A	*HP*	NS
F26	M	Comp het	16	72108203	CCT	C	*HP*	FS
F26	M	Comp het	3	65456156	T	A	*MAGI1*	NS
F26	M	Comp het	3	65456154	A	T	*MAGI1*	NS
F26	M	Comp het	7	151945071	G	GT	*MLL3*	FS
F26	M	Comp het	7	151874050	T	G	*MLL3*	NS
F26	M	Hemi	X	135593322	A	G	*HTATSF1*	NS
F26	M	Hemi	X	149931185	G	A	*MTMR1*	NS
F26	M	Hom	9	112900341	G	GGAAGCT	*PALM2-AKAP2*	NS
F26	M	Hemi	X	15474123	G	T	*PIR*	NS

F26	M	Hom	18	42456670	C	CTCTT	*SETBP1*	FS
F27	F	Comp het	7	131865473	G	A	*PLXNA4*	NS
F27	F	Comp het	7	131883311	C	T	*PLXNA4*	NS
F27	F	Hom	6	109850199	AAC	A	*AKD1*	FS
F27	F	Hom	15	40268998	G	GGACGAC	*EIF2AK4*	NS
F27	F	Hom	17	46608184	G	GGGGCGCTGT	*HOXB1*	NS
F27	F	Hom	7	15725797	ATGG	A	*MEOX2*	NS
F27	F	Hom	1	3566625	AGCAGGCTG	A	*WRAP73*	INDEL
F28	F	Comp het	20	52773992	C	T	*CYP24A1*	NS
F28	F	Comp het	20	52788189	C	T	*CYP24A1*	NS
F28	F	Comp het	10	47000019	G	A	*GPRIN2*	NS
F28	F	Comp het	10	46999596	G	A	*GPRIN2*	NS
F28	F	Comp het	4	123179882	T	G	*KIAA1109*	NS
F28	F	Comp het	4	123207867	T	G	*KIAA1109*	NS
F28	F	Comp het	16	84514205	G	A	*KIAA1609*	NS
F28	F	Comp het	16	84516214	G	A	*KIAA1609*	NS
F28	F	Comp het	1	17084536	TGGAACA	T	*MST1P9*	NS
F28	F	Comp het	1	17085427	T	C	*MST1P9*	NS
F28	F	Comp het	1	17087582	G	A	*MST1P9*	NS
F28	F	Comp het	17	70845790	G	A	*SLC39A11*	NS
F28	F	Comp het	17	70944008	C	T	*SLC39A11*	NS
F28	F	Hom	2	170632960	C	CA,CAA	*KLHL23*	FS
F28	F	Hom	5	112824048	T	TGCC	*MCC*	NS
F28	F	Hom	7	15725797	ATGG	A	*MEOX2*	NS
F28	F	Hom	7	131241029	GGGCGAC	G	*PODXL*	NS
F28	F	Hom	8	10467652	G	C	*RP1L1*	NS
F28	F	Hom	18	42456670	C	CTCTT	*SETBP1*	FS
F28	F	Hom	3	42251577	C	CGGA	*TRAK1*	NS
F28	F	Hom	13	100622667	TGGC	T	*ZIC5*	NS
F28	F	Hom	7	99662511	GTAGT	G	*ZNF3*	FS
F29	F	Comp het	7	48313854	G	A	*ABCA13*	NS
F29	F	Comp het	7	48312484	A	G	*ABCA13*	NS
F29	F	Comp het	3	182923984	G	A	*MCF2L2*	NS
F29	F	Comp het	3	183097166	G	A	*MCF2L2*	NS
F29	F	Comp het	19	54314254	C	T	*NLRP12*	NS
F29	F	Comp het	19	54301638	G	C	*NLRP12*	NS
F29	F	Comp het	8	101721932	CT	C	*PABPC1*	FS
F29	F	Comp het	8	101719201	A	G	*PABPC1*	NS
F29	F	Comp het	2	179582913	C	T	*TTN*	NS
F29	F	Comp het	2	179454969	G	A	*TTN*	NS
F29	F	Comp het	20	57766294	C	G	*ZNF831*	NS
F29	F	Comp het	20	57769291	C	T	*ZNF831*	NS
F29	F	Hom	12	103352171	C	CGCA	*ASCL1*	NS
F29	F	Hom	4	1389005	T	C	*CRIPAK*	NS
F29	F	Hom	6	32006214	CCTG	C	*CYP21A2*	NS
F29	F	Hom	19	54746081	T	C	*LILRA6*	NS
F29	F	Hom	7	100635591	C	A	*MUC12*	NS
F29	F	Hom	4	147560457	T	TGGC	*POU4F2*	NS
F29	F	Hom	19	43708978	TC	T	*PSG4*	FS
F29	F	Hom	2	179486037	C	A	*TTN*	NS
F29	F	Hom	2	179396782	C	G	*TTN*	NS
F29	F	Hom	19	44589999	TCTC	T	*ZNF284*	NS

Appendix 2: Inherited recessive and X-linked SNPs and indels that pass filters, in fetuses with structural abnormalities (preliminary round of analysis).
CQ = consequence of mutation; FS= frameshift coding; NS = non-synonymous. F31-F33 were sequenced subsequent to these analyses.

ID	SEX	GT	CHR	POS	REF	ALT	Gene	CQ
F1	M	Hemi	X	129146962	C	T	*BCORL1*	NS
F1	M	Hemi	X	34149726	G	A	*FAM47A*	NS
F1	M	Comp het	11	93806297	C	G	*HEPHL1*	NS
F1	M	Comp het	11	93808384	A	G	*HEPHL1*	NS
F1	M	Hemi	X	108868195	C	A	*KCNE1L*	STOP
F1	M	Hemi	X	151869653	A	G	*MAGEA6*	NS
F1	M	Comp het	8	48719844	G	A	*PRKDC*	NS
F1	M	Comp het	8	48848319	C	A	*PRKDC*	NS
F1	M	Hemi	X	117960384	G	A	*ZCCHC12*	NS
F1	M	Comp het	19	12358606	G	A	*ZNF44*	NS
F1	M	Comp het	19	12383893	G	A	*ZNF44*	NS
F2	F	Comp het	19	49113161	C	T	*FAM83E*	NS
F2	F	Comp het	19	49113215	G	A	*FAM83E*	NS
F2	F	Comp het	4	37445867	C	T	*KIAA1239*	NS
F2	F	Comp het	4	37446545	C	A	*KIAA1239*	NS
F2	F	Comp het	20	36868106	G	A	*KIAA1755*	NS
F2	F	Comp het	20	36870301	C	T	*KIAA1755*	NS
F2	F	Comp het	20	60886088	C	T	*LAMA5*	NS
F2	F	Comp het	20	60892813	G	A	*LAMA5*	NS
F2	F	Comp het	1	222802423	G	A	*MIA3*	NS
F2	F	Comp het	1	222802652	T	C	*MIA3*	NS
F3	M	Comp het	16	1470583	C	G	*C16orf91*	NS
F3	M	Comp het	16	1476330	T	C	*C16orf91*	NS
F3	M	Comp het	9	90500202	A	G	*C9orf79*	NS
F3	M	Comp het	9	90502542	T	C	*C9orf79*	NS
F3	M	Comp het	17	20799179	C	G	*CCDC144NL*	NS
F3	M	Comp het	17	20799281	G	A	*CCDC144NL*	NS
F3	M	Hemi	X	49104709	C	T	*CCDC22*	NS
F3	M	Comp het	6	138752868	C	A	*NHSL1*	NS
F3	M	Comp het	6	138794490	G	A	*NHSL1*	NS
F3	M	Hemi	X	9863050	C	T	*SHROOM2*	NS
F5	M	Comp het	8	12957657	C	T	*DLC1*	NS
F5	M	Comp het	8	13356860	G	C	*DLC1*	NS
F5	M	Hemi	X	119394834	G	A	*FAM70A*	NS
F5	M	Hemi	X	31089928	G	A	*FTHL17*	NS
F5	M	Hemi	X	135430934	C	A	*GPR112*	NS
F5	M	Hemi	X	99551442	G	C	*PCDH19*	NS
F5	M	Hemi	X	114426292	G	A	*RBMXL3*	NS
F5	M	Comp het	2	179430460	A	G	*TTN*	NS
F5	M	Comp het	2	179497758	A	G	*TTN*	NS
F5	M	Comp het	2	179579172	C	T	*TTN*	NS
F5	M	Hemi	X	117528073	A	C	*WDR44*	NS
F6	F	Hom	19	41754430	G	A	*AXL*	NS
F6	F	Comp het	7	30818142	T	G	*FAM188B*	NS
F6	F	Comp het	7	30825544	C	A	*FAM188B*	NS
F6	F	Comp het	7	103141235	G	A	*RELN*	NS
F6	F	Comp het	7	103205827	G	C	*RELN*	NS
F6	F	Comp het	1	8418331	C	T	*RERE*	NS
F6	F	Comp het	1	8418909	C	T	*RERE*	NS
F7	F	Comp het	19	9018166	A	G	*MUC16*	NS
F7	F	Comp het	19	9082960	G	A	*MUC16*	NS
F7	F	Comp het	13	45148705	TTGC	T	*TSC22D1*	NS
F7	F	Comp het	13	45149973	G	A	*TSC22D1*	NS
F7	F	Comp het	2	179399071	G	A	*TTN*	NS
F7	F	Comp het	2	179431633	C	T	*TTN*	NS
F7	F	Comp het	2	179641112	C	A	*TTN*	NS
F8	M	Comp het	2	160688217	T	C	*LY75-CD302*	NS
F8	M	Comp het	2	160738803	G	A	*LY75-CD302*	NS
F8	M	Hemi	X	153040414	C	T	*PLXNB3*	STOP
F8	M	Hemi	X	16870553	C	T	*RBBP7*	NS
F8	M	Hemi	X	99920314	G	T	*SRPX2*	NS
F8	M	Comp het	2	179399677	C	T	*TTN*	NS
F8	M	Comp het	2	179412829	C	T	*TTN*	NS
F8	M	Comp het	2	179591953	C	G	*TTN*	NS
F8	M	Comp het	16	74937918	C	T	*WDR59*	NS
F8	M	Comp het	16	74990380	G	A	*WDR59*	NS
F9	M	Comp het	7	48318614	T	G	*ABCA13*	NS
F9	M	Comp het	7	48547481	C	T	*ABCA13*	NS

F9	M	Hemi	X	152814163	G	A	*ATP2B3*	NS
F9	M	Hemi	X	49103316	G	A	*CCDC22*	NS
F9	M	Comp het	3	130300740	C	T	*COL6A6*	STOP
F9	M	Comp het	3	130381038	G	A	*COL6A6*	NS
F9	M	Comp het	20	57428948	G	C	*GNAS*	NS
F9	M	Comp het	20	57430118	C	G	*GNAS*	NS
F9	M	Comp het	10	30315676	G	A	*KIAA1462*	NS
F9	M	Comp het	10	30316500	A	ACTG	*KIAA1462*	NS
F9	M	Comp het	7	100679024	A	G	*MUC17*	NS
F9	M	Comp het	7	100685477	A	T	*MUC17*	NS
F9	M	Comp het	16	2806466	C	T	*SRRM2*	NS
F9	M	Comp het	16	2817604	G	A	*SRRM2*	NS
F9	M	Comp het	16	2817749	C	G	*SRRM2*	NS
F9	M	Comp het	2	234878910	C	T	*TRPM8*	NS
F9	M	Comp het	2	234891850	G	A	*TRPM8*	NS
F10	F	Comp het	19	3546264	C	T	*C19orf28*	NS
F10	F	Comp het	19	3551120	TC	T	*C19orf28*	FS
F10	F	Comp het	10	85955337	C	A	*CDHR1*	NS
F10	F	Comp het	10	85961593	C	T	*CDHR1*	NS
F10	F	Comp het	12	124330648	G	A	*DNAH10*	NS
F10	F	Comp het	12	124413109	T	C	*DNAH10*	NS
F10	F	Comp het	1	39851427	G	A	*MACF1*	NS
F10	F	Comp het	1	39901245	A	G	*MACF1*	NS
F11	M	Hemi	X	71521598	G	A	*CITED1*	NS
F11	M	Hemi	X	3242339	T	C	*MXRA5*	NS
F11	M	Hemi	X	30322699	T	C	*NR0B1*	NS
F11	M	Comp het	4	57777171	C	G	*REST*	NS
F11	M	Comp het	4	57796913	C	T	*REST*	NS
F12	F	Comp het	20	29631562	A	G	*FRG1B*	NS
F12	F	Comp het	20	29631580	A	G	*FRG1B*	NS
F12	F	Comp het	2	179396782	C	G	*TTN*	NS
F12	F	Comp het	2	179414177	G	A	*TTN*	NS
F12	F	Comp het	2	179484593	C	T	*TTN*	NS
F12	F	Comp het	2	179486037	C	A	*TTN*	NS
F12	F	Comp het	2	179549707	G	A	*TTN*	NS
F12	F	Comp het	2	179599473	C	G	*TTN*	NS
F12	F	Comp het	6	56999585	C	A	*ZNF451*	NS
F12	F	Comp het	6	57012673	C	T	*ZNF451*	NS
F13	M	Hemi	X	110980029	G	C	*ALG13*	NS
F13	M	Hemi	X	134713929	C	G	*DDX26B*	NS
F13	M	Comp het	4	79238620	C	T	*FRAS1*	NS
F13	M	Comp het	4	79353746	C	A	*FRAS1*	NS
F13	M	Hemi	X	135314244	G	A	*MAP7D3*	NS
F13	M	Comp het	15	42145586	G	A	*SPTBN5*	NS
F13	M	Comp het	15	42154034	C	T	*SPTBN5*	NS
F13	M	Hemi	X	12904292	T	A	*TLR7*	NS
F13	M	Comp het	2	1459885	A	G	*TPO*	NS
F13	M	Comp het	2	1544464	C	T	*TPO*	NS
F14	F	Comp het	20	49508015	T	C	*ADNP*	NS
F14	F	Comp het	20	49508508	C	T	*ADNP*	NS
F14	F	Comp het	2	242144345	T	G	*ANO7*	NS
F14	F	Comp het	2	242162665	C	T	*ANO7*	NS
F14	F	Comp het	1	214818291	G	A	*CENPF*	NS
F14	F	Comp het	1	214819026	A	C	*CENPF*	NS
F14	F	Comp het	6	46660511	T	A	*TDRD6*	NS
F14	F	Comp het	6	46661479	G	T	*TDRD6*	NS
F15	F	Comp het	5	148586585	A	G	*ABLIM3*	NS
F15	F	Comp het	5	148627397	C	T	*ABLIM3*	NS
F15	F	Comp het	5	82833426	A	G	*VCAN*	NS
F15	F	Comp het	5	82835589	T	C	*VCAN*	NS
F16	M	Comp het	16	1470583	C	G	*C16orf91*	NS
F16	M	Comp het	16	1476330	T	C	*C16orf91*	NS
F16	M	Comp het	9	90500202	A	G	*C9orf79*	NS
F16	M	Comp het	9	90502542	T	C	*C9orf79*	NS
F16	M	Comp het	17	20799179	C	G	*CCDC144NL*	NS
F16	M	Comp het	17	20799281	G	A	*CCDC144NL*	NS
F16	M	Comp het	6	138752868	C	A	*NHSL1*	NS
F16	M	Comp het	6	138794490	G	A	*NHSL1*	NS
F16	M	Hemi	X	9863050	C	T	*SHROOM2*	NS

F17	F	Comp het	16	2369688	A	T	*ABCA3*	NS
F17	F	Comp het	16	2374481	T	C	*ABCA3*	NS
F17	F	Comp het	13	42875878	C	T	*AKAP11*	NS
F17	F	Comp het	13	42876835	T	G	*AKAP11*	NS
F17	F	Comp het	1	68960131	T	C	*DEPDC1*	NS
F17	F	Comp het	1	68960186	T	C	*DEPDC1*	NS
F17	F	Comp het	1	26303228	G	A	*PAFAH2*	NS
F17	F	Comp het	1	26317303	C	T	*PAFAH2*	NS
F17	F	Comp het	7	75070377	T	A	*POM121C*	NS
F17	F	Comp het	7	75070840	C	A	*POM121C*	NS
F18	M	Hemi	X	105855323	T	C	*CXorf57*	NS
F18	M	Hemi	X	44703940	A	G	*DUSP21*	NS
F18	M	Hemi	X	138644189	C	T	*F9*	NS
F18	M	Hemi	X	55650995	C	T	*FOXR2*	NS
F18	M	Hemi	X	131842557	G	C	*HS6ST2*	NS
F18	M	Hemi	X	119077233	C	G	*NKAP*	NS
F18	M	Comp het	3	135969390	A	C	*PCCB*	NS
F18	M	Comp het	3	136019898	C	T	*PCCB*	NS
F18	M	Hemi	X	129546514	G	A	*RBMX2*	NS
F18	M	Comp het	2	179452447	T	C	*TTN*	NS
F18	M	Comp het	2	179611552	C	T	*TTN*	NS
F18	M	Comp het	16	72831357	C	CTTGTTG	*ZFHX3*	NS
F18	M	Comp het	16	72831629	G	A	*ZFHX3*	NS
F18	M	Comp het	16	72832550	A	C	*ZFHX3*	NS
F19	M	Hom	16	84229207	C	T	*ADAD2*	NS
F19	M	Comp het	14	105415079	G	T	*AHNAK2*	NS
F19	M	Comp het	14	105416541	C	G	*AHNAK2*	NS
F19	M	Comp het	20	61326565	C	T	*C20orf90*	NS
F19	M	Comp het	20	61331818	C	G	*C20orf90*	NS
F19	M	Comp het	12	7521535	A	G	*CD163L1*	NS
F19	M	Comp het	12	7527284	C	T	*CD163L1*	NS
F19	M	Hemi	X	107431191	T	A	*COL4A6*	NS
F19	M	Comp het	3	52409413	C	G	*DNAH1*	NS
F19	M	Comp het	3	52426643	G	A	*DNAH1*	NS
F19	M	Comp het	5	13759007	G	A	*DNAH5*	NS
F19	M	Comp het	5	13883075	C	T	*DNAH5*	NS
F19	M	Comp het	2	84880481	C	G	*DNAH6*	NS
F19	M	Comp het	2	84924743	C	G	*DNAH6*	NS
F19	M	Comp het	5	132534965	G	A	*FSTL4*	NS
F19	M	Comp het	5	132939589	C	T	*FSTL4*	NS
F19	M	Hemi	X	2793951	T	C	*GYG2*	NS
F19	M	Hom	21	47831802	C	T	*PCNT*	NS
F19	M	Comp het	16	71712805	C	A	*PHLPP2*	NS
F19	M	Comp het	16	71724598	T	C	*PHLPP2*	NS
F19	M	Hemi	X	152225801	G	A	*PNMA3*	NS
F19	M	Hemi	X	84362764	G	A	*SATL1*	NS
F19	M	Hemi	X	9863131	G	A	*SHROOM2*	NS
F20	M	Comp het	8	61769198	C	G	*CHD7*	NS
F20	M	Comp het	8	61778448	C	T	*CHD7*	NS
F20	M	Comp het	6	131247845	T	A	*EPB41L2*	NS
F20	M	Comp het	6	131277390	G	A	*EPB41L2*	NS
F20	M	Hemi	X	152860096	C	T	*FAM58A*	NS
F20	M	Comp het	5	90002053	A	G	*GPR98*	NS
F20	M	Comp het	5	90059270	C	A	*GPR98*	NS
F20	M	Hemi	X	154290176	C	G	*MTCP1NB*	NS
F20	M	Hemi	X	153697736	C	T	*PLXNA3*	NS
F20	M	Hemi	X	153716622	T	C	*SLC10A3*	NS
F20	M	Comp het	1	12337460	T	A	*VPS13D*	NS
F20	M	Comp het	1	12337667	C	T	*VPS13D*	NS
F20	M	Comp het	1	12378274	C	T	*VPS13D*	NS
F21	M	Hemi	X	100911707	G	A	*ARMCX2*	NS
F21	M	Comp het	16	1265315	G	A	*CACNA1H*	NS
F21	M	Comp het	16	1270350	G	A	*CACNA1H*	NS
F21	M	Hemi	X	65824281	G	A	*EDA2R*	NS
F21	M	Hemi	X	135593768	G	A	*HTATSF1*	NS
F21	M	Hom	14	104641986	C	G	*KIF26A*	NS
F21	M	Hemi	X	135303057	T	C	*MAP7D3*	NS
F21	M	Hemi	X	63490871	TC	T	*MTMR8*	FS
F21	M	Hemi	X	3239828	T	C	*MXRA5*	NS

F21	M	Comp het	6	51656129	C	G	*PKHD1*	NS
F21	M	Comp het	6	51768399	A	T	*PKHD1*	NS
F22	M	Comp het	8	91049129	C	G	*DECR1*	NS
F22	M	Comp het	8	91057198	A	G	*DECR1*	NS
F22	M	Comp het	15	45411495	C	A	*DUOXA1*	NS
F22	M	Comp het	15	45412435	G	A	*DUOXA1*	NS
F22	M	Hemi	X	135314244	G	A	*MAP7D3*	NS
F22	M	Comp het	2	152346522	G	A	*NEB*	NS
F22	M	Comp het	2	152384078	C	T	*NEB*	NS
F22	M	Hom	5	140553876	T	C	*PCDHB7*	NS
F22	M	Comp het	15	62212467	C	T	*VPS13C*	NS
F22	M	Comp het	15	62212770	T	C	*VPS13C*	NS
F23	M	Comp het	1	170952626	T	C	*C1orf129*	NS
F23	M	Comp het	1	170961328	C	T	*C1orf129*	NS
F23	M	Hom	5	74018232	A	G	*GFM2*	NS
F23	M	Hemi	X	19398315	C	T	*MAP3K15*	NS
F23	M	Hemi	X	135313855	T	C	*MAP7D3*	NS
F23	M	Comp het	11	70332311	C	T	*SHANK2*	NS
F23	M	Comp het	11	70336479	C	T	*SHANK2*	NS
F23	M	Comp het	2	179404498	G	C	*TTN*	NS
F23	M	Comp het	2	179424272	C	A	*TTN*	NS
F23	M	Comp het	2	179454530	C	T	*TTN*	NS
F23	M	Comp het	2	179610967	C	T	*TTN*	NS
F25	M	Hemi	X	39932564	C	T	*BCOR*	NS
F25	M	Comp het	1	22150156	G	T	*HSPG2*	NS
F25	M	Comp het	1	22206977	C	T	*HSPG2*	NS
F25	M	Comp het	1	156497776	C	CA	*IQGAP3*	FS
F25	M	Comp het	1	156504308	G	A	*IQGAP3*	NS
F25	M	Hemi	X	102755132	TC	T	*RAB40A*	FS
F25	M	Hemi	X	132160102	G	A	*USP26*	NS
F26	M	Comp het	1	145515394	A	G	*GNRHR2*	NS
F26	M	Comp het	1	145515696	A	T	*GNRHR2*	NS
F26	M	Hemi	X	135593322	A	G	*HTATSF1*	NS
F26	M	Hemi	X	149931185	G	A	*MTMR1*	NS
F26	M	Hemi	X	15474123	G	T	*PIR*	NS
F28	F	Comp het	20	52773992	C	T	*CYP24A1*	NS
F28	F	Comp het	20	52788189	C	T	*CYP24A1*	NS
F28	F	Comp het	4	123179882	T	G	*KIAA1109*	NS
F28	F	Comp het	4	123207867	T	G	*KIAA1109*	NS
F28	F	Comp het	16	84514205	G	A	*KIAA1609*	NS
F28	F	Comp het	16	84516214	G	A	*KIAA1609*	NS
F28	F	Comp het	17	70845790	G	A	*SLC39A11*	NS
F28	F	Comp het	17	70944008	C	T	*SLC39A11*	NS
F29	F	Comp het	7	48312484	A	G	*ABCA13*	NS
F29	F	Comp het	7	48313854	G	A	*ABCA13*	NS
F29	F	Comp het	3	182923984	G	A	*MCF2L2*	NS
F29	F	Comp het	3	183097166	G	A	*MCF2L2*	NS
F29	F	Comp het	19	54301638	G	C	*NLRP12*	NS
F29	F	Comp het	19	54314254	C	T	*NLRP12*	NS
F29	F	Comp het	7	75052435	C	T	*POM121C*	NS
F29	F	Comp het	7	75070334	C	T	*POM121C*	NS
F29	F	Hom	2	179396782	C	G	*TTN*	NS
F29	F	Comp het	2	179454969	G	A	*TTN*	NS
F29	F	Hom	2	179486037	C	A	*TTN*	NS
F29	F	Comp het	2	179582913	C	T	*TTN*	NS
F29	F	Comp het	20	57766294	C	G	*ZNF831*	NS
F29	F	Comp het	20	57769291	C	T	*ZNF831*	NS
F31	F	Comp het	15	80452844	G	A	*FAH*	NS
F31	F	Comp het	15	80464527	C	A	*FAH*	NS
F33	F	Comp het	1	981151	T	C	*AGRN*	NS
F33	F	Comp het	1	985378	G	A	*AGRN*	NS
F33	F	Comp het	19	33183575	T	A	*NUDT19*	NS
F33	F	Comp het	19	33200127	T	C	*NUDT19*	NS

Appendix 3: High-quality, rare, coding, inherited recessive and X-linked SNPs and indels (final round of analysis).
GT = genotype; CQ = consequence of mutation; FS= frameshift coding; NS = non-synonymous.