Identification of Novel Genes Involved in Brain Abnormality Syndromes Using Next Generation Sequencing



Malcolm Freeman Howard

Oriel College

University of Oxford

A thesis submitted for the degree of

Master of Science by Research

In Clinical Medicine

July 2014

Identification of Novel Genes Involved in Brain Abnormality Syndromes Using Next Generation Sequencing

Malcolm Freeman Howard
Oriel College, University of Oxford, MRes, July 2014

ABSTRACT

Since the 1980s there has been a steady increase in the number of genes identified which play a role in human disease. The Human Genome Project and arrival of Next Generation Sequencing (NGS) have helped accelerate the rate of discovery such that the recent update of the Human Gene Mutation Database (HGMD) lists over 140 thousand genetic variants linked to human disease [1-3]. However, causative genes have only been identified in half of the 7000 rare monogenic disorders [4,5]. It is predicted that within 10 years the majority of disease-causing genes will be identified [4]. In many ways the era of next-generation sequencing has become the "gold rush" of disease gene discovery.

The aim of this thesis was to demonstrate the validity of an NGS approach to identify novel disease genes associated with brain abnormality phenotypes. In a cohort of 23 families with a range of brain abnormality phenotypes, novel disease genes were found in 2 families and there was an overall clinical diagnosis rate of 38%. The results from this thesis have already expanded the spectrum of brain abnormality mechanisms to include the gene *PGAP3* [6]. The strategy outlined will become part of the clinical framework for the diagnosis and identification of brain abnormality syndromes with NGS.

Abstract 2
Table of Contents 3
List of Figures 5
List of Tables 6
Acknowledgements 7
Publications 8
Abbreviations Used 9
Web Resources 10

Chapter 1: Introduction 12
1.1. Traditional approach of disease gene identification 14
1.2. Next-generation sequencing 16
1.3. Brain abnormalities 23

Chapter 2: Methods and Materials 28
2.1. Recruitment of patients 28
2.2. Extraction of DNA 30
2.3. Quantification and QC of DNA aliquot 30
2.4. SNP arrays for homozygosity mapping 31
2.4.1. Analysis of array data for copy number variations and allelic imbalances in Nexus 33
2.4.2. Mapping of shared homozygous (autozygous) regions using SNP array data 34
2.4.3. Calculation of inbreeding coefficient 36
2.5. Whole Exome Sequencing 37
2.5.1. Mapping of exome sequencing reads 38
2.5.2. Merging of lanes and removal of duplicate reads 41
2.5.3. Exome genotype concordance with SNP array data 42
2.5.4. Exome contamination and merging check 44
2.5.5. Exome Read Count Examination 45
2.6. Variant calling with Platypus 46
2.6.1. Family and trio based variant calling 48
2.6.2. Exome Region file creation 49
2.7. Ingenuity Variant Analysis 51
2.7.1. Modification of VCF files for Ingenuity Variant Analysis 52
2.7.2. Ingenuity VCF import and annotation pipeline 53
2.7.3. Default Ingenuity Variant Analysis filtering 54
2.7.4. Validation of Ingenuity Variant Analysis 58
2.7.5. Custom IVA filtering developed for variants associated with brain abnormality phenotypes 58
2.7.6. Incidental findings 61
2.8. Candidate variant examination in IGV 62
2.9. PCR primer design for candidate variants 63
2.9.1. PCR amplification of candidate variant regions 67
2.10. Sanger sequencing of candidate variant regions 68
2.10.1. Analysis of Sanger sequencing data 70
2.11. Sequencing of *PGAP3* 71

Chapter 3: Results 73
3.1. Bioinformatics and variant calling pipeline 73
3.1.1. Comparison of autozygous region mapping with PLINK and HomozygosityMapper 73

3.1.2. Comparison of estimation of Coefficient of Inbreeding with reported degree of consanguinity 75
3.1.3. Effect of the reference genome on variant calling 76
3.1.4. Effect of duplicate reads on Platypus variant calling 78
3.1.5. Exome concordance analysis 80
3.1.6. Exome contamination and lane merging analysis 81
3.1.7. Exome read count examination 83
3.1.8. Platypus variant caller 91
3.1.9. Validation of Ingenuity Variant Analysis for filtering and analysing sequence data 99
3.2. Identification of candidate variants associated with brain abnormality phenotypes 105
3.2.1. Overview of cohort 105
3.2.2. Clinically relevant brain abnormality findings for patients and their families 108
3.2.3. Variants in Glycosylphophatidylinositol (GPI) pathway gene *PGAP3* implicated in brain abnormality phenotype for BRC_41 and siblings 113
3.2.4. Identification of *PI4KA* variants in polymicrogyria family (1139E1, 1139F1, BRC_231) 130
3.2.5. Identification of *SOX11* variant in BRC_146 139

Chapter 4: Discussion 142
4.1. Development of bioinformatics pipeline 142
4.1.1. Ingenuity Variant Analysis 143
4.1.2. Inclusion of the mitochondrial genome 145
4.1.3. Platypus minimum flank settings and variant annotation 145
4.1.4. Variant Databases 146
4.2. Clinical application of NGS 147
4.2.1. Comparative sensitivity of Sanger sequencing and NGS 147
4.2.2. Previous clinical testing of known genes 148
4.2.3. Clinically relevant findings and the benefit of genetic diagnosis 150
4.2.4. Reverse phenotyping 151
4.3. Novel genes for brain abnormality phenotypes 151
4.3.1. PGAP3 152
4.3.2. PI4KA 153
4.4. Future 155

References **157**
Appendix Table 1 **172-73**
Appendix Table 2 **174**
Reprint of Paper Published in Am J Hum Gen **175**

List of Figures

Chapter 2

Figure 2.2: Example of advantage of trio variant calling with Platypus. 49
Figure 2.3: Example of modified exome target regions 50
Figure 2.4: Filtering cascade and filtering options in IVA 55
Figure 2.5 Examples of biological context filtering in IVA 57
Figure 2.6 Likely sequencing artefacts due to poly A nucleotide repeat 63
Figure 2.7: Primer design in UCSC Genome Browser. 64

Chapter 3

Figure 3.1: Shared homozygous chr11 regions of BRC_179 and BRC_180 from HomozygosityMapper 75
Figure 3.2: Examples of calls on chromosome X unique to Human37 in two unrelated samples. 78
Figure 3.3: Histogram plot of freemix values from verifyBamID programme of exome samples. 82
Figure 3.4: Comparison of freemix values between the TruSeq and two batches run with NimbleGen kit. 83
Figure 3.5: Mean target read depth for each chromosomes for samples run with SeqCap (NimbleGen) capture kit. 85
Figure 3.6: IGV visualization of BRC_153 coverage compared with parental samples. . 88
Figure 3.7: Example of the two common variants in *ASPM* shown in IGV and in UCSC Genome Browser. 92
Figure 3.8: Plot showing the distribution of variation between mF=0, mF=10, and Platypus 0.2.0 VCFs. 95
Figure 3.9: WGS and WES comparison of mitochondrial regions visualized in IGV. 97
Figure 3.10: Pedigree of BRC_41 114
Figure 3.11: Shared region of homozygosity in BRC_41 and affected siblings. 116
Figure 3.12: Ingenuity variant analysis filtering cascade for BRC_41 117
Figure 3.13: Phylogenetic conservation of *PGAP3* variant p.G92D 118
Figure 3.14: Sanger validation of PGAP3 variants in BRC_41 and family 118
Figure 3.15: Representation of growing GPI molecule and approximate positions of mutation in *PGAP3* 124
Figure 3.16: Flow cytometric analysis for function of *PGAP3* variant in BRC_41 126
Figure 3.17: Photographs of *PGAP3* affected individuals. 129
Figure 3.19: Phylogenetic Conservation of *PI4KA* variant p.D1854N 131
Figure 3.20: Sanger validation of *PI4KA* mutations 133
Figure 3.21: Comparison of Sanger traces of *PI4KA* variant p.D1854N with new primer 134
Figure 3.22: Comparison of BLAT results from sequences using old and re-designed primers. 135
Figure 3.24: Sanger validation of SOX11 variant in BRC_146 139

List of Tables

Chapter 2

Table 2.1: Summary of patients recruited and experimental approach 29
Table 2.2: PLINK commands used for calculation of shared ROH 36
Table 2.3: Sequences of 100 bp forward and reverse Illumina reads from FASTQ..... 39
Table 2.4: Command line for alignment and mapping 41
Table 2.5: Command line operations of for QC steps 44
Table 2.6: Platypus variant caller command line settings..... 48
Table 2.7: PCR and Sanger sequencing primers for candidate variants 66
Table 2.8: Forward and reverse *PGAP3* PCR and Sanger sequencing primers..... 71

Chapter 3

Table 3.1: Concordance of shared homozygous regions between HomozygosityMapper and PLINK 74
Table 3.2: Comparison of estimation of coefficient of inbreeding with reported degree of consanguinity 76
Table 3.3: Comparison of total number of variants called in samples mapped to Human37 or hs37d5..... 77
Table 3.4 Comparison of total number of variants called in samples with and without duplicate reads. 79
Table 3.6 Validation of python script compared with output from QualiMap v0.8.1. 84
Table 3.7: Mean and median depth of coverage for 43 samples run with NimbleGen capture kit and percentage covered at 10x, 20x and 30x. 86
Table 3.8: Mean and median coverage of three samples run with TruSeq capture kit and percentage covered at 10x, 20x and 30x. 87
Table 3.9: Read counts of failed parental samples BRC_153F and BRC_153M 87
Table 3.10: Brain abnormality genes with less than 10X coverage in samples. 89
Table 3.12: Comparison of variant calls with different minFlank settings. 93
Table 3.13: Number of concordant variants called in mitochondrial regions in WGS compared with WES. 98
Table 3.14: Comparison of Perl and Python script modified genotype field outputs. 104
Table 3.15: Candidate brain abnormality variants identified by WES 106
Table 3.16 Example read pair used for manual alignment with UCSC BLAT tool. 136

Acknowledgements

I would like to thank my supervisors Dr Jenny Taylor and Sir John Bell for their help, encouragement, guidance and support. I am extremely grateful to Dr Alistair Pagnamenta who supervised me in the laboratory and taught me many of the techniques described in this project. I would also like to thank Dr Usha Kini who is the primary clinician that I worked with in this thesis, who recruited the patients into the study and has been very supportive of the research.

Lastly, I would like to thank Erika and our son Gavin for being amazing and supportive over this last year.

Publications

Howard, Malcolm F., et al. "Mutations in *PGAP3* Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation." The American Journal of Human Genetics 94.2 (2014): 278-287.

Abbreviations used

aCGH	array comparative genomic hybridization
ACMG	American College of Medical Genetics
AD	allele depth
ALP	Alkaline Phosphatase
BAM	binary sequence alignment file
bp	base pair
BWA	burrows-wheeler aligner
CCDS	Consensus NCBI coding sequence
CHO	Chinese hamster ovary
cnLOH	copy neutral loss of heterozygosity
CNV	copy number variation
CSS	Coffin-Siris syndrome
DGV	database of genomic variation
DP	depth
EFMR	epilepsy limited to females with intellectual disabilities
EIEE	Early Infantile Epileptic Encephalopathy (Ohtahara)
ER	endoplasmic reticulum
FTP	file transfer protocol
GPI	Glycosylphosphatidylinositol
GPI-AP	GPI anchored protein
HGMD	Human Gene Mutation Database
HM	HomozygosityMapper
IGV	Integrative Genomics Viewer
IVA	Ingenuity Variant Analysis
MAF	Minimum Allele Frequency
Mb	mega-base (million base pairs)
mF	minimum flank setting in Platypus
mtDNA	Mitochondrial Genome
NGS	next-generation sequencing
NR	number of reads
NV	number of variant reads
OFC	occipitofrontal head circumference
OMIM	Online Mendelian Inheritance in Man Database
OUH	Oxford University Hospitals
PCR	polymerase chain reaction
PGD	pre-implantation genetic diagnosis
PMG	polymicrogyria
PPC	parent-parent-child
RGL	region genetics laboratory (Oxford)
ROH	region of homozygosity
SNP	single nucleotide polymorphism

SNV	single nucleotide variat
VCF	variant call format file
VEP	variant effect predictor
VF	variant frequency
WES	whole exome sequencing
WGS	whole genome sequencing
WGS500	Wellcome Trust Centre for Human Genetics whole genome sequencing project (274 WGS of mixed ancestry)
WTCHG	Wellcome Trust Centre for Human Genetics

Web Resources

1000 Genomes	http://www.1000genomes.org/
Burrows-Wheeler Aligner	http://bio-bwa.sourceforge.net/
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
ClustalW	http://www.clustal.org/clustal2/
Coriell Cell Repositories	http://ccr.coriell.org/
dbSNP	http://www.ncbi.nlm.nih.gov/snp/
HomozygosityMapper	http://www.homozygositymapper.org/
Ingenuity Variant Analysis	https://variants.ingenuity.com/
NHLBI Exome Sequencing Project (ESP) Exome Variant Server	http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man (OMIM)	http://www.omim.org
Oxford Brain Abnormality Research Group	http://www.brainabnormalities.org.uk
Platypus	http://www.well.ox.ac.uk/platypus
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
SIFT	http://sift.jcvi.org/
UCSC Genome Browser	http://www.genome.ucsc.edu

Chapter 1: Introduction

Chapter 1: Introduction

The identification of genetic mutations responsible for congenital brain malformations is a challenging process due to the rarity and heterogeneous nature of these conditions [7]. Studies have shown that there are over 13,000 genes expressed in the human brain [8,9]. The sheer biological complexity of human brain development means that for many neurological conditions, the underlying cause remains uncertain. Discovery of novel brain abnormality disease genes provides insight into brain development and disease mechanisms. A greater understanding of the mechanisms underlying many brain abnormalities will improve patient care and treatment [7].

The arrival of next-generation sequencing (NGS) has brought with it a new era of disease gene discovery. These technologies have also confirmed the considerable genetic variability that exists between any two healthy individuals [10]. The interpretation of this variation is a vital step in the process of disease gene discovery. In a healthy individual there are typically 2 million single nucleotide variants (SNV) and within those 2 million SNVs, there can be between 20 to 60 thousand novel variants [11]. It is apparent that even when considering a healthy individual, it is a difficult task to interpret the sheer amount of genetic variation present. When the task is to find the variant responsible for causing human disease among all of the variation, it is a considerable undertaking.

There are several strategies that make the identification of disease-causing variants easier. Analysis can be focused on the exome, the 1% of the genome responsible for protein

coding, where roughly 85% of the known genetic causes for Mendelian diseases have been found [5,12]. Candidate region approaches, such as linkage analysis and homozygosity mapping, can be used to narrow down the region that potentially harbours the disease genes. Another approach is to sequence several affected individuals from the same family, and examine the shared variants segregating with the disease. In the case of sporadic dominant conditions, parent-parent-child trios can be sequenced and the data can be filtered for novel, *de novo* variants only present in the child. All of these strategies have been used with varying degrees of success to identify Mendelian disease-causing variation [13]. The difficult and complex task that geneticists and clinicians now face is how to apply NGS technologies in a clinical setting where there is a direct impact on patients.

There is considerable evidence that NGS can be used clinically as a diagnostic tool to identify the molecular cause of a genetic disorder [14-17]. The difficulty is the interpretation and analysis of the data produced and the large number of putative pathogenic variants that may be identified [11]. The process is relatively straightforward when a previously reported variant is found in a known disease-causing gene that matches the patient's phenotype. The literature supporting the variant is examined, which can be time consuming, but provided it is sound, it can support a clinical diagnosis for the patient. When a novel variant is found in a known disease-causing gene, the process becomes more complex. The effect on protein coding, relative to known variants and protein domains, needs to be assessed. In addition, there must be some persuasive evidence that the gene product could, if abnormal, lead to a particular disease phenotype.

Based on this evidence, the variant could be classified as pathogenic. In some cases, functional laboratory studies are required to assess the *in vitro* impact of the variant on the protein before an assessment can be made. The identification of variants in a novel disease gene requires substantial evidence and laboratory work before they can truly be considered pathogenic in a clinical setting.

1.1. Traditional approach of disease gene identification

The method of dideoxynucleotide (di-deoxy) terminator sequencing, Sanger sequencing was developed by Sanger and colleagues in 1977 [18]. The advent of the Polymerase Chain Reaction (PCR) in 1983 was one of the first technologies that significantly increased the pace of genetic research [19,20]. The PCR reaction made it possible to rapidly amplify DNA fragments, which greatly increased the efficiency of Sanger sequencing. Over the next 25 years, Sanger sequencing was the main tool used by geneticists in determining DNA sequences of suspected disease genes.

The traditional approach to disease gene identification was Sanger sequencing of candidate genes usually in regions of genetic linkage. The first step in linkage analysis is to collect pedigrees of families in which the disease is segregating. Subsequent analysis involves genotyping polymorphic DNA markers across the genome in these pedigrees. By comparing the distribution of genetic markers in affected and unaffected individuals, the location of the disease gene in the genome can be narrowed. In certain cases the locus of the disease gene can be further narrowed by cytogenetic rearrangements involving the

gene [21]. Once the location is identified, the process involves capturing and sequencing genetic segments expressed from that region. All of the genes isolated by this method must be sequenced and analysed for mutations in affected individuals to determine if any are consistent with causing the disorder [22]. This particular step is labour intensive and in some cases can take several years before the actual disease-causing gene is isolated and cloned [23]. This positional cloning strategy was one of the major approaches for disease gene discovery. Positional cloning also offered one of the first approaches for finding disease genes where no prior knowledge of the biochemical impact of the genetic defect was required [21]. In 1986, positional cloning was used to identify disease genes responsible for Chronic Granulomatous Disease and Duchenne Muscular Dystrophy [24,25]. During the same period, the *RB1* gene associated with Retinoblastomas was cloned [26]. The identification of these disease genes was followed by the discovery of the Cystic Fibrosis gene in 1989 [27,28].

The principal of homozygosity mapping was proposed in 1987 as an extension of linkage analysis to locate disease genes in consanguineous children with Mendelian disorders [29]. The regions of homozygosity (ROH) were predicted to be the likely location of autosomal recessive disease genes. An advantage of this approach was that large pedigrees of affected individuals did not need to be collected to identify candidate regions. In recent years, the resolution of disease gene mapping has improved with the switch from microsatellite repeats to high density Single Nucleotide Polymorphism (SNP) based arrays. As well as refining the results of linkage analysis, these arrays can implicate novel disease genes by helping identify rare copy number variation (CNV) [5].

Linkage studies and positional cloning were a traditional method of finding the genomic region responsible for a disease and eventually the gene. This approach was limited by the rarity of these conditions and the limited number of large pedigrees with multiple affected family members. If the candidate region contained many genes, it could take years to find the specific gene responsible for the disorder. In the case of Microcephalin, *MCPH1*, a gene associated with autosomal recessive microcephaly, homozygosity mapping narrowed down the disease locus to a 5.3 million bp (Mb) region on chromosome 8 [30]. It was not until four years later in 2002 that the gene was identified [31]. Despite these limitations, by the year 2003, variations of the approach for identifying genes with linkage and positional cloning identified over 1200 Mendelian disease genes [32].

1.2. Next-generation sequencing

A limiting factor of the positional cloning in linkage region approach was always the number of genes identified in each region. Unless there was an obvious candidate, there could be dozens, if not hundreds of genes to Sanger sequence. The release of the first commercial high throughput genetic sequencing platform, the 454 system in 2005 brought with it a rapid method of sequencing thousands of genes which revolutionized the process of disease gene identification [33].

Chapter 1: Introduction

In 2009, the cost of sequencing a whole genome was still over one hundred thousand dollars which was prohibitive for large-scale disease gene studies [5,34]. Since the cost of whole genome sequencing (WGS) was so expensive and the majority of variants associated with Mendelian disease were found in the coding regions of the genome [5,12], a method was developed that only targeted the exome for sequencing. The exome sequencing method involved the capture and enrichment of DNA from exonic regions followed by NGS. The proof of principle of exome sequencing was the sequencing of 12 whole exomes and the identification of variants in the known disease gene *MYH3* in two patients with Freeman-Sheldon Syndrome [35].

The advent of whole exome sequencing (WES) was a technological breakthrough for gene discovery using NGS. WES of four affected individuals with Miller syndrome found variants in the gene *DHODH,* which encodes a key enzyme in pyrimidine biosynthesis [36,37]. The result was reinforced by finding variants in *DHODH* in other families with Miller Syndrome using Sanger sequencing [37]. This was the first successful application of NGS to discover the gene responsible for a rare Mendelian disease of unknown cause. Concurrent studies were also being done with WGS. In 2010, whole genome sequencing was used to identify the novel disease gene *SH3TC2* in the heterogeneous condition Charcot-Marie-Tooth Disease [38] and in another study, *PTPN11* as causing the dominant condition metachondromatosis [39]. Despite the feasibility of WGS, it was the inexpensive sequencing of the exome that dominated NGS studies over the next four years.

One of the problems with the traditional disease gene approaches was in finding the gene responsible for sporadic dominant conditions. These conditions were usually so severe that the responsible genetic variant would not remain in the population due to selective pressures and lack of reproductive fitness [40]. Since these conditions were very rare and did not follow standard inheritance models, it was not possible to use linkage analysis in large families or genome wide association studies to find the disease gene locus [41]. For many of these conditions it was possible that a *de novo* mutation was responsible for the disease, not an inherited variant. The normal mutation rate in the human germline has been estimated at 1.18×10^{-8} per position, which roughly equates to 74 novel SNVs per genome per generation [36,41-43]. This mutation rate has been found in several studies to roughly correlate to one non-synonymous mutation per generation [43-46]. If the mutation were to occur in a well-conserved and biologically vital gene, it could result in a severe phenotype.

NGS platforms also allow for the screening of thousands of genes for the presence of *de novo* point mutations. The first example of *de novo* variants that cause disease was in Schinzel-Giedon syndrome [47]. The exomes of four unrelated affected individuals were sequenced and all were found to have heterozygous *de novo* mutations in *SETBP1*. The *de novo* origin of the variants was confirmed by Sanger sequencing *SETBP1* in the parents. In a similar approach, exome sequencing identified mutations in *MLL2* as a cause of Kabuki syndrome from a group of 12 unrelated probands [48]. Based on the nature of *de novo* variants, it was hypothesized that sequencing parent-parent-child trios with suspected *de novo* variants could, in many cases, identify the cause of idiopathic mental

retardation [45]. The trio approach involved WES of the proband and their parents. All of the variants in the proband that followed Mendelian inheritance would also be present in the parents. Thus the *de novo* variant or variants could be found by filtering for variants in the proband that were not present in the parents. The proof of this approach was in a study where WES identified likely pathogenic *de novo* variants in six out of ten individuals with unexplained learning disability [45]. The *de novo* trio approach was quickly adopted and shown to find potentially pathogenic variants in sporadic conditions such as autism [49] and schizophrenia [50]. Larger studies suggest that likely causative *de novo* mutations can be identified in 16-43% of cases with sporadic learning disability [51,52].

In addition to the *de novo* approach, there have been several other exome sequencing strategies that were effective in finding new disease genes. Many of these strategies were based on traditional approaches such as linkage analysis and homozygosity mapping. A linkage study utilizing WES data identified *PIGV* as a cause of Hyperphosphatasia with Mental Retardation Syndrome [53]. Homozygosity mapping using only exome sequencing data found rare variants in *SERPINF1* in patients with autosomal-recessive Osteogenesis Imperfecta [54]. The homozygosity mapping approach has been particularly effective in finding the cause of recessive cognitive disorders in consanguineous families [55]. In one of the largest studies of its kind to date, next generation sequencing of 136 consanguineous families with autosomal recessive learning disability identified variants in 23 known genes and implicated an additional 50 novel candidate genes [55].

The power of the NGS approach is the ability to rapidly sequence all of those genes identified in the regions of interest. WES and WGS also provide an unbiased approach, which can eliminate other potential causative genes. A limitation of homozygosity mapping can be the assumption that the variant is in one of the homozygous regions. It is always possible that there are compound heterozygous variants that explain the phenotype and the consanguinity is irrelevant. By sequencing all of the genes in parallel, the homozygous regions can be screened and the rest of the coding regions can also be checked for deleterious variants in candidate genes.

Early WES studies started to show the potential clinical applications of NGS. The ability to quickly diagnose a genetic condition can have a dramatic impact on a patient and their family [56]. The discovery of variants causing Mendelian disorders can lead to diagnosis, improved screening [57] and potentially treatment [58]. An interesting example of WES in diagnosing a genetic condition was for a patient with suspected Bartter syndrome [59]. WES identified a variant in *SLC26A3*, which had previously been associated with congenital chloride diarrhoea and can be treated with salt replacement therapy. The diagnosis demonstrated the potential clinical utility of WES. Another early application of WES in a hospital led to a lifesaving diagnosis and treatment of a young child who had intractable bowel disorder. WES identified a deleterious variant in the known disease gene *XIAP* [15]. This diagnosis led to a subsequent allogenic haematopoeitic progenitor cell transplant and apparent successful treatment of the disease. In both examples, the phenotype of the patient was not indicative of the disorders associated with those genes so neither had been screened by Sanger sequencing. This example highlights the ability

of NGS to screen all genes in an unbiased manner. As was the case here, genetic diseases can have variable phenotypes depending on the nature of the variant. It is also possible that the patients are clinically misdiagnosed and NGS offers an unbiased method to provide a genetic diagnosis. The large-scale clinical utility of WES has recently been demonstrated by the identification of the genetic mutation in 25% of 255 consecutive patients referred for evaluation of possible Mendelian genetic disorders [14].

There are also examples where known disease genes have been screened and missed by Sanger sequencing. An example was the diagnosis of neonatal diabetes mellitus using NGS which identified a missense variant in *ABCC8* which had been previously missed by Sanger sequencing [60]. It had an immediate clinical impact as the patient was treated with oral sulfonylurea instead of insulin therapy. When the Sanger sequencing data from the initial screening was reviewed, the variant was present but had obviously been missed. The bioinformatics workflow that is associated with NGS avoids these types of errors, as heterozygous variants, even those at low frequency can be detected and reported in the VCF files.

NGS has allowed scientists to study the genetic cause of a range of inherited diseases at an unparalleled rate. After the initial WES proof of concept studies there has been considerable adoption of the method for rare disease research. In the first two years, WES led to the identification of over 30 new genes in Mendelian diseases [13]. By December 2012, 178 novel genes had been discovered with WES [4]. The rate of novel disease gene discovery is only expected to continue with the increasing affordability of WGS.

WES has been the method of choice for NGS studies, since the cost per sample is significantly lower than WGS. The price to sequence a human genome, however, has steadily decreased from the first genome in 2001[2,3] and is now at the long sought after goal of the thousand dollar genome [34]. In January 2014, Illumina declared that the goal had been reached with their new sequence platform, the HiSeq X Ten [61]. Increasing affordability suggests that WGS will soon become the preferred tool for rare disease research. Although the majority of Mendelian variants have been found in the coding regions, it is possible that this could be slightly inflated since most studies have focused on the exome [5]. The likely increase of WGS highlights an on-going issue with NGS research, the interpretation of large quantities of data. The challenge of traditional disease gene identification was finding and narrowing the region that harboured the disease gene. The challenge of NGS is the interpretation of thousands of variants which can be a labour-intensive process.

There are several well-established methods by which NGS data can be interpreted and analysed. The variant annotation, filtering and ranking are all time consuming and in some cases complex. To manually review the literature of every single novel variant present in the data would be a significant and daunting undertaking [62]. Variant annotation packages and software can be used to facilitate the analysis; the difficulty is integrating all of these tools into a pipeline. The annotation must include a diverse set of data and needs to be regularly updated. Many of the tools are based in the UNIX command line, which can be a challenge for laboratory-based scientists. There is also an

unmet clinical need for a pipeline that is automatable, reproducible and auditable. Current practices involve analysis of NGS data with custom scripts and a variety of programmes. Once candidates are identified, these are provided to the clinical scientist who based on the phenotype of the patient, assess the likely pathogenicity of the candidates. A pipeline that would provide data in a usable format directly to the clinical scientist to independently analyse would be of benefit. The clinical scientist would be able to use their expertise in the phenotypes to directly assess the deleterious variants present in the data.

The aim of this thesis is to show how NGS analytical methods, combined with new bioinformatics tools, can be used to find novel brain abnormality genes and efficiently identify variants in known genes using WES. The bioinformatics work described in this thesis is an efficient approach in which to move from raw sequence data to candidate variants. This thesis describes the development of a bioinformatics pipeline which builds on the WTCHG pipeline and begins to address bioinformatics needs for the clinical diagnostic application of WES, thus creating the basis for similar WGS studies.

1.3. Brain abnormalities

In this thesis, the term brain abnormalities encompasses a range of phenotypes. These include structural brain malformations and neurodevelopmental disorders. All patients have developmental delay with a range of phenotypes including microcephaly, seizures

and polymicrogyria (PMG). A broad group of patients was included in the study because this highlights the utility of the research approach in analysing a complex phenotype.

Chromosomal aberrations are the most common known cause of developmental delay [63]. In the past, the first line clinical test for children presenting with brain abnormalities not characteristic of a known condition was conventional karyotyping [64]. These cytogenetic methods were used to detect disorders resulting from large chromosomal imbalances but the diagnostic yield was only roughly ten percent in individuals with neurodevelopmental disorders [65]. This method was limited by the resolution of light microscopy. Array comparative genomic hybridization (aCGH) is now the first line test as it can detect large and small chromosomal aberrations. Despite improvements in methods and resolution, the diagnostic rate is still less than 20% [64,66,67]. In cases that do not present with clear pathogenic chromosomal aberration, it is possible that small point mutations or deletions in genes critical for proper development are responsible for the phenotype.

The heterogeneous nature of many brain abnormalities makes it quite difficult to identify the responsible genetic variants [7]. In the case of microcephaly, there are over 400 entries in the Online Mendelian Inheritance in Man Database (OMIM) that list this phenotype in the clinical synopsis [7,68]. Many of these conditions are associated with deletion and duplication syndromes but there are a significant number of genes with specific variants associated with microcephaly. For autosomal recessive primary microcephaly, there are currently twelve genes directly implicated with the phenotype

[69]. To screen all these genes by traditional Sanger sequencing is a lengthy process. Although targeted NGS is an option, WES has advantages over panel testing. The NGS approach described in this thesis allows for many candidate genes to be screened in an unbiased manner. If no variants are found in known genes, it is possible that the cause of the phenotype is a mutation in a novel disease gene as was the case with the identification of *WDR62* as a novel disease gene for microcephaly with WES [70]. NGS provides an efficient method for screening all known genes as well as an opportunity to discover novel ones.

Another complex brain abnormality phenotype is epilepsy. The molecular basis of epilepsies and seizures are difficult to identify when they involve a brain abnormality phenotype. In the case of neuronal migration disorders such as PMG and lissencephaly, the aberrant cortical developmental is hypothesized to be responsible for the seizure phenotype [71]. There is also significant phenotypic variability; in the microcephaly gene *WDR62* there are reported cases of patients presenting with and without seizures [70,72]. In glycosylphosphatidylinositol (GPI) anchor related disorders, seizures are often reported as part of the brain abnormality phenotype. In these disorders essential neuronal proteins, such as *TNAP* are no longer anchored to lipid cell surface which could be responsible for the epilepsy [73]. It is also possible that there are multi-genic effects, which could account for the complex phenotype. A large WES study identified four patients with pathogenic variants in multiple genes resulting in novel multi-genic phenotypes [14]. One of the patients presented with a brain abnormality phenotype including seizures, congenital abnormalities and mental retardation. Exome sequencing of

the patient identified causative variants in *SCN2A* and *POMT2,* which together explained the complex phenotype [14]. Traditional approaches of disease gene discovery and classification would have focused on a monogenic diagnosis to explain the phenotype. Only through NGS, where all genes are screened, can compound phenotypes such as this be explained.

For many neurological conditions, the underlying cause still remains uncertain. Since seizures and brain abnormalities exhibit extensive genetic heterogeneity, the NGS approach of this project is suited to screen for known brain abnormality genes and to potentially discover novel disease genes. This thesis demonstrates the power of these approaches for clarifying the aetiology of these diseases and providing patients and their families with a diagnosis.

Chapter 2: Methods and Materials

Chapter 2: Methods and Materials

2.1. Recruitment of patients

For this research project, 23 patients with brain abnormalities and 25 unaffected family members were recruited via the OUH Clinical Genetics Department from the on-going Structural Brain Abnormalities and Learning Disability study with ethics approval from the Wales Research Ethics Committee (12/WA/0001) (Table 2.1). The study has recruited more than 200 subjects for genetics research with one or more of the following inclusion criteria: abnormal occipitofrontal head circumference (OFC), dysmorphic cranial feature not characteristic of known congenital disorders and severe learning disability.

Patients with phenotypes and clinical history not suggestive of known disorders were selected for this exome sequencing project following a negative screen for pathogenic CNVs and where appropriate, clinical testing for variants in *ASPM*, the most common cause of autosomal recessive primary microcephaly [74]. A next-generation sequencing (NGS) approach with whole exome sequencing (WES) was undertaken to identify the cause of the patient's brain abnormality phenotype. For most families, a trio-based approach was undertaken (Table 2.1). When consanguinity was reported (or suspected), an autozygosity mapping strategy with Single Nucleotide Polymorphism (SNP) arrays and WES was considered as an alternative.

Table 2.1: Summary of patients recruited and experimental approach
All of the patients recruited for this research are listed in the table along with the experimental approach undertaken. Large families are grouped together. When only one family member was WES, the proband is marked with a *.

Patient ID	Gender	Father	Mother	Patient Phenotype Summary
Trio sequencing approach where proband and both parents were WES				
BRC_115	Female	BRC_115F	BRC_115M	Relative microcephaly, seizures and significant developmental delay
BRC_16	Female	BRC_16F	BRC_16M	Microcephaly (-5SD), cerebellar hemisphere hypoplasia and speech issues.
BRC_59	Female	BRC_59F	BRC_59M	Microcephaly (-5SD to -6SD), seizures, and hypoplastic cerebellum and pons.
BRC_81	Male	BRC_81F	BRC_81M	-5 to -6SD microcephaly, cortical dysplasia
BRC_90	Female	BRC_90F	BRC_90M	Microcephaly and a simplified cortex
BRC_91	Male	BRC_91F	BRC_91M	Microcephaly, short stature, and learning difficulties.
BRC_146	Male	BRC_146F	BRC_146M	Microcephaly, trigonocephaly and cerebellar vermis atrophy.
BRC_153	Female	BRC_153F	BRC_153M	Angelman syndrome like, microcephaly, seizures.
BRC_225	Female	BRC_225F	BRC_225M	Lissencephaly with contractures and microcephaly
BRC_230	Female	BRC_230F	BRC_230M	Severe microcephaly (-5 SD)
Homozygosity Mapping where nuclear family was run on SNP arrays and single proband (*) underwent WES				
BRC_41*	Male	BRC_41F	BRC_41M	Microcephaly, seizures and hyperphosphatasia
BRC_42	Female			
BRC_216	Female			
BRC50*	Female	BRC_50F	BRC_50M	Microcephaly and cerebellar atrophy.
BRC102*	Male	BRC_102F	BRC_102M	Microcephaly
BRC103	Female			
ZUR	Male			UNAFFECTED SIBLING
BRC_180*	Male	BRC_180F	BRC_180M	Microcephaly
BRC_179	Female			
Consanguineous families where single proband (*) was WES				
BRC_224*	Female	BRC_224F	BRC_224M	Microcephaly and seizures.
BRC_226*	Male	BRC_226F	BRC_226M	Microcephaly, seizures and hypocalcaemia.
Large families with multiple affected children/fetuses where entire family was WES				
1139E1	Male	BRC219	BRC220	Polymicrogyria, cerebellar hypoplasia and bilateral talipes
1138F1	Female			
BRC_231	Female			
BRC_83	Male	BRC_83F	BRC_84M	Microcephaly.
BRC_84	Male			
Single proband (*) was WES (hyperphosphatasia / GPI disorder suspected)				
BRC_228*	Male	unavailable	unavailable	Microcephaly, seizures and raised ALP. Adopted child.
Patient with previous clinical diagnosis of 22q11 deletion syndrome (Proband and Mother WES)				
BRC_232	Male	Unavailable	BRC_232M	Child with 22q11 and polymicrogyria.

2.2. Extraction of DNA

Blood samples were collected from patients during a visit to a National Health Service clinic by a trained health care professional. 2 to 5ml of blood was collected in ethylenediaminetetraacetic acid (EDTA) tubes. The DNA was extracted in the Oxford Regional Genetics Laboratory (RGL) at the Churchill Hospital using the Gentra Autopure LS method (Qiagen) [75]. This is a robotic system which automates the well-established liquid phase purification Gentra Puregene method. Following extraction, the DNA was re-suspended in 100µl of re-hydration solution buffer Tris EDTA (TE) and normalized to 100ng/µl. After extraction the samples were stored at -80°C until required. The samples and patient data were anonymized for this research project. Aliquots of patient DNA samples were prepared and verified by RGL Clinical Scientists before being provided for the project.

2.3. Quantification and QC of DNA aliquot

The quality of the DNA samples provided for the research study was assessed with a Nanodrop 1000 (Thermo Scientific). The absorption ratio 260/280nm and 260/230nm UV light was used to assess the purity of the DNA sample [76]. A 260/280nm absorption ratio of ~1.8 to 2.0 was used as the standard for pure DNA. The ratio of 260/230nm was used as the secondary measure of purity, with expected values in the range of 2.0-2.2. DNA samples were quantified for the SNP arrays and Polymerase Chain Reaction (PCR) experiments with the Nanodrop and normalized to 50ng/µl and 20ng/µl respectively. The Nanodrop estimates the quantity of DNA based on the amount of light absorbed by the sample and the known optical density of double-stranded DNA. Since spectrophotometry

does not directly measure DNA molecules there could be a tendency to overestimate the quantity of DNA when impurities are present [77]. Therefore to ensure accurate sample preparation for exome sequencing, the Qubit Fluorometer (Invitrogen) was used to quantify and normalize DNA samples to 50ng/µl before sequencing. The Qubit system was generally established to have greater accuracy and precision when quantifying DNA compared against spectrophotometry [78]. The technology was based on the use of dyes which fluoresce when bound to DNA. The total amount of fluorescent signal is proportional to the quantity and DNA which can be calculated when compared against known standards [78].

2.4. SNP arrays for homozygosity mapping

In this project, samples from consanguineous families were run on CytoSNP12v2 (Illumina) arrays in order to identify regions of homozygosity which might harbour candidate genes. The array data was also used to check for CNVs that may have been missed by prior screens, estimate the coefficient of inbreeding and for genotype concordance with exome sequencing data. In one family, SNP arrays were used as an initial screen for CNVs.

The 3 day experimental protocol and conditions specified by the manufacturer were followed [79]. In brief, the protocol for each sample was as follows: 200ng of patient DNA was denatured and neutralized followed by an overnight genomic amplification step which increased the total quantity of DNA by several thousand fold. Samples were end-point fragmented, isopropanol precipitated and collected by centrifugation. Precipitated

DNA was re-suspended and hybridized to the BeadChip in an overnight step. Subsequent wash steps removed un-hybridized and non-specific hybridized DNA from the BeadChip. The BeadChip was dried in a vacuum desiccator before the intensities of the beads' fluorescence was assayed using the Iscan system (Illumina). SNP arrays are based on the principle that the DNA fragments from the sample anneal to specific oligo sequences covalently bound to one of the 300,000 beads on the array. The oligo hybridizes to the DNA fragment in the region adjacent to the target SNP. The nucleotide from the extension step is incorporated into the oligo based on the target SNP position on the template DNA fragment. The nucleotide is specific to the allele of the SNP from the DNA fragment and is designed to fluoresce in the red or green colour spectrum. The Iscan system reads the BeadChip by taking high-resolution images of the light emitted from the laser excited fluorescent nucleotides bound to the individual beads. Illumina software calculates the log-R value or copy number, based on the total florescence and the B-allele frequency based on the relative fluorescence of red or green at each bead position.

The data from the Iscan system was imported into Genome Studio v-2011.1 (Illumina) along with the DMAP file containing the specific bead locations and manifest file which described the SNP and probe content of the array. SNPs were genotyped with the standard cluster file HumanCytoSNP-12v2-1-H. The software determined the log-R and B-allele frequency for every SNP based on the established cluster position information present in the cluster file. The overall SNP call rate was used as the quality metric for each sample and was calculated by the software as the total percentage of SNPs and

probes genotyped on the array. A call rate of greater than 99% was expected for good quality samples [80]. The array data was inspected visually, chromosome by chromosome, in the Illumina Genome Viewer package as another check of the quality of the data, to ensure that the data was not noisy and to inspect for large aberrations.

2.4.1. Analysis of array data for copy number variations and allelic imbalances in Nexus

The SNP array data from this project was analysed in Nexus Copy Number software (BioDiscovery), since it has algorithms for calling structural variation and regions of copy neutral loss of heterozygosity (cnLOH). It also has better tools for visualization and downstream data analysis. Although all of the patient's DNA samples had previously been screened for CNVs, the SNP array data was still scrutinized for any CNVs that might have been missed by the previous testing. The data was exported from Genome Studio as a final report with the calculated B-allele frequencies and log-R values for each probe and imported into Nexus version 7. Copy number variants and allelic events were called using the SNPRank Segmentation algorithm which is a validated and accurate method for calling CNVs [81]. In this study, the specific parameters for the algorithm were modified from the defaults to improve the stringency for CNV and allelic event calls. The significance threshold was increased from 5.0×10^{-7} to 1.0×10^{-8}, requiring regions to have greater statistical difference from the normal to be called. The minimum number of contiguous probes for a region was increased from 3 to 4 and the thresholds for a gain and a loss were increased, which is simply the log-R value at which an event was called. These settings were based on in-house experience for accurate and validated CNV and cnLOH calls with Illumina arrays and Nexus [82-84].

A list of CNVs and allelic imbalances was generated for each sample. The regions called by the Nexus algorithm were assessed visually to check the validity each call. CNVs in regions that appeared noisy with sets of probes above and below the call threshold in series were discarded. Those aberrations that appeared to be real were evaluated based on the frequency in the Database of Genomic Variants (DGV) [85]. Nexus plotted the most recent release of DGV against the array data and reported if any of the CNV calls overlapped with known variants. If the CNV was in DGV and not reported in the literature, it was deemed to be benign. CNVs that encompassed coding regions were scrutinized for genes that had previously been implicated in brain abnormalities or similar disorders. The final criterion for evaluating previously unreported CNVs was the segregation in family, if it was inherited from an unaffected parent or present in an unaffected sibling it was assessed as non-causative.

2.4.2. Mapping of shared homozygous (autozygous) regions using SNP array data

In the patients where consanguinity was reported an autozygous mapping approach was taken to find shared regions of homozygosity. After the SNP array data was imported into Nexus, the SNPRank algorithm was used to generate a list of allelic imbalances. Those genomic segments identified as being regions of cnLOH were compared visually between affected family members. The borders of each region were inspected closely to locate the first heterozygous SNP call. Regions close to the centromere and other regions with a low probe density that appeared consistently in unrelated samples were considered array artefacts and ignored. Successive neighbouring cnLOH regions were examined to

ascertain if they were one contiguous region. In cases where there were only a small number of heterozygous probes separating large cnLOH regions, the clustering of each SNP was inspected in Genome Studio. If the SNPs appeared on the clustering map apart from other samples and close to the border of the homozygous call, it was assumed that the SNP in question was likely homozygous. For SNPs in neighbouring regions separated by a small number of potentially miscalled SNPs, the whole region was interpreted as homozygous.

The autozygous regions identified visually in the software were further confirmed with the online tool HomozygosityMapper (HM) [86] and PLINK [87]. HM was used as it calculates shared ROHs, links with the GeneDistiller tool and also has an online interface for visualizing the data. The genotype information was exported from Genome Studio with the Illumina Genotype calls, AA, AB, BB, and uploaded directly to HM website. The tool utilizes custom Perl functions, which screen the data for blocks of homozygous genotypes in contiguous regions. The end of regions were plotted when a heterozygous SNP was found which was not flanked by 7 or more homozygous genotypes on each side [88]. Once the shared homozygous regions were plotted, the GeneDistiller tool [89] was used to search the regions for candidate genes based on patient phenotypic terms listed in Online Mendelian Inheritance in Man (OMIM) [90] gene reports.

Since HM is a relatively new tool, the regions were also calculated using the UNIX command line based programme PLINK v1.90b. The SNP array data was exported from Genome Studio using Illumina's PLINK Input Report Plug-in which generated the

necessary PLINK input files. The pedigree files were annotated to include the familial relationships and the affected status. The shared homozygous regions were calculated using the --homozyg --homozyg-group options that outputted a list of shared homozygous regions (Table 2.2). To identify homozygous regions, PLINK scanned the genotype data for regions of extended homozygosity of a minimum length, allowing for occasional missed genotypes and heterozygote calls [87].

Table 2.2: PLINK commands used for calculation of shared ROH
UNIX command line operations used for running PLINK and calculating shared ROH

Programme	Version	Command line
PLINK	1.90b1a	plink --file *sample* --homozyg --homozyg-group

2.4.3. Calculation of inbreeding coefficient

Since several of the patients were from consanguineous families, an estimation of the degree of consanguinity was calculated, as it is important to check that the reported family pedigree matches the amount of homozygosity found in the genome. The inbreeding coefficients were estimated as the total sum of homozygous regions above a specific length divided by the total length of the autosomal genome covered by the cytoSNP-12 arrays. The total length of the autosomal genome used for these calculations was 2,836,047,788 bp based on the hg19 human reference genome [91]. The total length of homozygous regions was calculated based on the consensus homozygous regions calculated by Nexus and PLINK ignoring homozygous centromere regions present in all samples. A 4 Mb minimal region was used as the cut-off for the calculation since homozygous regions up to that size are common in demonstrably outbred individuals

[92]. This method for estimating inbreeding coefficients has been shown to correlate strongly with inbreeding coefficients estimated from multigenerational pedigrees [92]. The coefficient of inbreeding could not be calculated with PLINK. PLINK calculates inbreeding coefficient based on the expected number of heterozygotes vs homozygous genotypes observed. For accurate inbreeding coefficients, the calculation requires allele frequencies of each SNP in an outbred population. The total cohort of cytoSNP-12 genotypes available was less than 80 samples, of which little was known about the family histories.

2.5. Whole Exome Sequencing

The patient samples were submitted to the WTCHG Core Sequencing facility for Whole Exome Sequencing on the HiSeq2500 or 2000. The target coverage for the exonic regions was 25-30x with over 80% of the exome covered at greater than 10x. This was based on in-house experience for variant identification [93] and the WTCHG Whole Genome Sequencing Project (WGS500) [94]. In most cases, the samples were run in multiplexed barcoded batches of 10 samples per lane on the HiSeq with 2 lanes run per sample group. An average yield of 150 to 200 million reads per lane of the HiSeq corresponds to approximately 15 to 20 million reads per sample depending on sample sequencing performance. The multiplexed two-lane approach was predicted to achieve the target coverage based on the Core's previous experience with exome sequencing. The exome capture kits used in this research study were the TruSeq Exome Enrichment kit (Illumina) and the SeqCap EZ Human Exome Library v3.0 (NimbleGen). The Core transitioned to the newer SeqCap exome kit during the study. The libraries were prepared from the

submitted samples according to the manufacturer protocol by the Core lab technicians. In brief, the libraries were prepared by fragmenting the DNA samples, followed by blunt ending and ligating of sequencing adaptors to the fragments. The exonic target regions were subsequently captured and the libraries were amplified before being sequenced on the HiSeq. The libraries were run in multiplexed sample groups on one lane of a HiSeq2500 or 2000 machine, using the 100bp paired-read format.

2.5.1. Mapping of exome sequencing reads

The raw sequencing data must be mapped to a reference genome for variant calling. Sequencing data produced for each sample from the Illumina HiSeq comprises two text files (FASTQ) containing a list of 100bp DNA sequences with associated quality scores. There is a corresponding DNA sequence in each text file representing the forward and reverse sequencing of the same molecule of DNA. The process of mapping involves finding the corresponding location in the reference genome that closely matches the sequences. The example shown in Table 2.3 can be mapped simply with the BLAT [95] tool which utilizes an index of the reference genome stored in the UCSC servers and finds the closest match to imputed sequence, chr19:52,130,749-52,130,949 (Figure 2.1). The process of mapping sequencing data is analogous but at a much larger scale involving millions of 100bp reads.

For this project, the data from each lane of the exome sequencing was run through the standard WTCHG bioinformatics pipeline for sequencing data to produce BAM files. BAM files are the binary version of Sequence Alignment Map (SAM) files. The pipeline

maps the raw sequence data, Illumina FASTQ files, to the Human Reference using the Burrows Wheeler Aligner (BWA) [96] in conjunction with Stampy [97].

Table 2.3: Sequences of 100 bp forward and reverse Illumina reads from FASTQ
Shown are two paired-end sequencing reads from a FASTQ file with the mapping quality scores removed.

@HISEQ2000-05:335:C38PFACXX:2:1113:6005:37843 2:N:0:CGATGT
CAATACGCCCCCTGACCTTGCAGCAGCAGGACAGAGCCGCTCTGGGACCCAT
AGATGTTCCAGGCCTTGCAGCTGACTTTGAGGTCGGAGCTGAGCCCCC
@HISEQ2000-05:335:C38PFACXX:2:1113:6005:37843 1:N:0:CGATGT
GACAGAGCCGCTCTGGGACCCATAGATGTTCCAGGCCTTGCAGCTGACTTTG
AGGTCGGAGCTGAGCCCCCCGTGGAGGATCAGGGAGCTGTTGGCCCAG

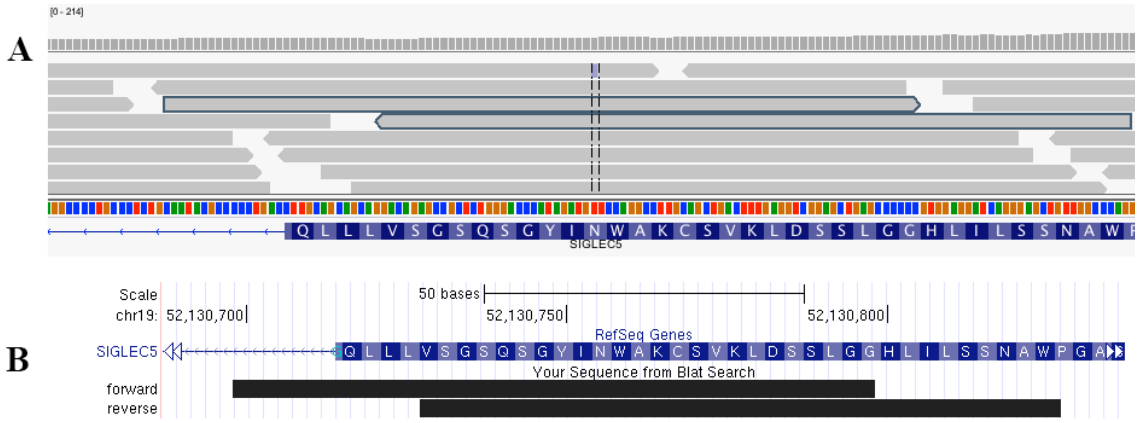


Figure 2.1: Forward and reverse read in IGV and UCSC Blat alignment
(A) 100bp WES sequencing reads are represented as grey bars in IGV. The reads from Table 1.1 are highlighted and align to exon 6 of *SIGLEC5*.
(B) The forward and reverse read pair (black) shown in UCSC Genome Browser. The reads were mapped by the BLAT tool and aligned to exon 6 of *SIGLEC5*.

The BWA aligner utilizes the Burrows-Wheeler transform algorithm [98] which is a computing resource-efficient method for aligning millions of reads to a reference genome. The algorithm is a text-based compression technique which functions in simple terms by rearranging character strings into runs of similar characters [98]. In the case of mapping sequencing reads, the compression technique allows for faster and efficient alignment of reads but tends to be less sensitive than the slower hash-based mapping

programmes [97]. In a manner analogous to the BLAT tool, hash-based programmes map reads by indexing the reads or reference genome into the computer memory. Since the entire reference or all of sequence reads are loaded into the computer memory, it is a slow and resource intensive process. Stampy utilizes a hash based table system to map reads with high sensitivity but can be run in a hybrid mode with BWA [97]. In this mode, BWA maps the majority of reads which closely match the reference and Stampy provides the sensitivity of a hash based programme for the difficult to map reads [97]. By mapping the sequencing data in a two-stage process with two complementary mapping programmes, the resultant data is produced in an accurate and computing resource efficient manner.

The BAM files produced by BWA and Stampy were sorted by coordinate order and indexed with the command line Picard [99] tool. The initial quality assessment of the sequencing and resultant BAM files was based on the QC statistics produced by the in-house and unpublished QC package pipeline-tools. All of the FASTQ and BAM files for each lane of data were downloaded from the WTCHG Core servers by File Transfer Protocol (FTP).

The initial cohort of exomes run on the SeqCap (NimbleGen) kit was mapped to the human GRCh37/hg19 reference. The subsequent exomes were mapped to the hs37d5 version of hg19 which includes a decoy chromosome. hs37d5 was developed by the Broad Institute and used in the final phase of the 1000 Genomes project. The hs37d5 reference genome contains an extra 35.4Mb of sequences which are supposed to attract mis-mapped reads and improve variant calls in chromosomal regions [100]. To maintain

consistency across family groups, the initial cohort of exomes was remapped to hs37d5. This was done in a manner similar to that described above, except I did it with the command lines shown in Table 2.4. The variant calls from 10 of the exomes mapped to both references were compared to ascertain the effect of the decoy chromosomes in this project.

Table 2.4: Command line for alignment and mapping
UNIX command line operations used for alignment, mapping, sorting, merging, duplicate removal and indexing of sequencing reads.

Programme	Version	Command line
BWA	0.7.5a	bwa aln -t 2 -q10 hs37d5.fa *sample_1,2*.fastq > *sample_1,2*.sai
BWA	0.7.5a	bwa sampe hs37d5.fa *sample_1*.sai *sample_1*.fastq *sample_2*.sai *sample_2*.fastq |
samtools	0.1.18	samtools view -b -o *sample*.bwa.bam -S -
Stampy	1.0.22	stampy.py -t 2 -g hs37d5 -h hs37d5 --bamkeepgoodreads --readgroup="`cat *sample*.rgArg.txt`" --comment=@*sample*.comments.txt -M *sample*.bwa.bam |
samtools	0.1.18	samtools view -bh -o *sample*.stampy.bam -S -
Picard	1.104	picard-Xmx4g SortSam.jar SO=coordinate I=*sample*.stampy.bam O=*sample*.bam
Picard	1.104	picard MergeSamFiles.jar I=*lane1*.bam I=*lane2*.bam
Picard	1.104	picard MarkDuplicates.jar VERBOSITY=ERROR REMOVE_DUPLICATES=true I=sample_dups.bam O=sample.bam M=sample.metrics.txt
Picard	1.104	BuildBamIndex.jar I=sample.bam O=sample.bam.bai

2.5.2. Merging of lanes and removal of duplicate reads

WES data was downloaded from the Core servers for each lane of sequencing as raw BAM files which required merging. In most cases, each exome sample was run on two lanes of the HiSeq producing two BAM files for each sample. The sample data sets were merged on the WTCHG Cluster server with the Picard command line tool (Table 2.4). Removal of duplicate reads was undertaken to decrease the number of false variants. PCR duplicates are identical DNA fragments that were amplified many times during the preparation of the sequencing library. If the artefacts over-represent a variant, it could cause the variant to be miscalled by the variant caller programme [101]. Duplicate reads were removed from the BAM using Picard. After merging and the removal of duplicate

reads from the BAM files, an index file was created with Picard (Table 2.4). The index file is a map or table of contents for the BAM file which allows for other programmes to efficiently access specific regions.

2.5.3. Exome genotype concordance with SNP array data

To assess the quality of the exome sequencing in this project the concordance of genotype calls from the SNP arrays was compared with that from the exome sequencing. Genotype calls from the CytoSNP12 arrays were determined with the Illumina Genome Studio software. The no-call threshold was set to 0.20, above the 0.15 standard suggested by Illumina for genotyping accuracy based on in-house experience that this level of stringency functions best for concordance analysis. The no-call threshold is based on the GenCall metric which assigns a score from 0 to 1 based on the distance a sample is from the centre of the cluster for that data point [102]. The further a SNP is from the cluster plot for that data point, the less likely the genotype call is to be accurate and the smaller the GenCall value.

The genotype data was exported as a tab-delimited text file from Genome Studio, including the SNP calls for each allele, the reported SNP for that position and the reported strand orientation. Intensity-only and zeroed SNPs were excluded from the report as they do not provide genotype information. The data was run though a Python script developed as part of this project which added a column for each SNP describing the hs37d5 reference allele, utilizing samtools faidx[103]. The script also read each genotype position in the file and checked that the SNP and allele call matched relative to the reference with the correct strandedness. Although Genome Studio was queried for the

genotype calls in the forward direction, it still had to be corrected for strand errors. The strand orientation of the allele calls was corrected by the script in any positions that were not reported on the positive strand.

To generate genotype calls from the exome data, Platypus was run in genotyping mode on each sample for which both WES and SNP array data existed (Table 2.6). A list of positions covered by the cytoSNP-12 array was turned into a template VCF file which was compressed and indexed and used as the variant list for Platypus's genotyping function. In genotyping mode, Platypus takes the list of SNPs in the template VCF and genotypes each position in the BAM file. The sample VCF files were split with intersectBED tool [104] (Table 2.5) into the regions targeted by NimbleGen V3 exome kit and those regions falling outside. This was done to assess the overall concordance inside and outside of the target regions.

The genotyped VCF file for the cytoSNP12 regions and the genotype call list from Genome studio were compared for overall genotype call concordance with another python script written for this project. The script functioned by reading each line of the VCF file, extrapolating a genotype based on reference and alternate allele for that position and the zygosity. The genotypes were sorted alphabetically to allow for string comparison with the second file, i.e. AG rather than GA. The genotype call from the exome VCF was then compared against the call from the array data. A file was created with all the non-concordant positions and the overall number of concordant and non-concordant positions. The overall concordance rate was calculated as the number of

concordant positions divided by the total number of SNPs present in the exome target region.

For the set of three exomes run with the TruSeq (Illumina) capture kit, a simple concordance check was done with Excel (Microsoft). A genotyped list of SNP positions was exported from Genome Studio as described above. The genotype calls were derived from the VCFs for those files with a simple IF function in excel. The two files were compared for concordance with the vLookup function. The overall concordance was calculated as the number of concordant positions in the exome VCF divided by the total number of genotyped positions. Genotype concordance with the array data was used as a metric to assess the performance of the exome sequencing.

Table 2.5: Command line operations of for QC steps
UNIX command line operations for QC assessment of exome concordance, contamination and read count.

Programme	Version	Command line
samtools	0.1.18	samtools faidx hs37d5.fa chr1:1-1
verifyBamID	1.1.0	verifyBamID.1.1.0 --vcf 1000G_v3_exome_regions.vcf --bam *sample*.bam --out *sample*.prefix --verbose
bedtools	2.17.0	intersectBed –a *sample*.vcf -b v3_regions.bed -v > *sample_outside*.vcf
bedtools	2.17.0	bedtools coverage -d -abam *sample*.bam -b coverage_check.bed > *sample*.coverage.txt

2.5.4. Exome contamination and merging check

Since SNP array data was only available for a subset of the exomes, the verifyBamID programme was used as a quality control metric for the exome data (Table 2.5). The programme is currently the standard exome QC method for the WTCHG Clinical Exome Project. The function is based on a mathematical model that relates observed sequence reads to hypothetical true genotypes based on 1000 Genomes allele frequencies [105]. The freemix metric, a value from 0 to 1 was used as the overall quality measure. The

validity of the metric was assessed by running verifyBamID on three simulated BAM files with differing degrees of contamination created by merging lanes of unrelated samples. Those samples with array data were also run through the programme, with the simulated genotype frequencies. The verifyBamID functioned to check that all of the merged BAM files were from the original sample. This was done firstly by checking for sample contamination and by also checking that merged BAM files came from the correct samples based on the original SM: (sample name) specified in the header file.

2.5.5. Exome Read Count Examination

The final quality control assessment for the sequencing was the overall read count in the target regions of the exome capture kits. The bedtools coverage tool [104] was run on all the merged BAM files for the exome data. The function was run from a script on the command line with the –d option to report the read depth at each base position in the every region from the BED file (Table 2.5). The read counts were only assessed for the target regions of the capture kit, including the mtDNA regions. The output from the function was run through a python script written for this project that flagged the target regions of the respective exome capture kits with little or no coverage. These regions were assessed to ensure that no significant genes of interest were missed with the exome sequencing. The list of brain abnormality genes was generated from the Ingenuity Knowledge Base (version 140408.002) (Appendix Table 1). The input BED file for bedtools coverage was annotated with a column that specified if a region included a gene of interest. The coverage of these specific genes was directly assessed with the python script. The output from the script included all of the genes of interest that had no coverage, less than 4X mean coverage and less than 10X mean coverage. The script also

reported the overall coverage on a per chromosome basis and the percentage of the target regions with greater than 10X, 20X and 30X coverage. To validate the output from the python script the coverage statistics were compared with an established javascript programme Qualimap v0.8.1[106] which also calculates the mean and median coverage.

2.6. Variant calling with Platypus

For this project a variant calling algorithm was required to call variants from the aligned sequencing data. Variant Call Format (VCF) files for each patient and family were created using the variant caller programme Platypus v0.5.2 [107] [108]. Platypus is a WTCHG developed tool, which has been established by several studies to accurately detect variants in high throughput sequencing data [82,93,94,109-111]. It was also the variant caller programme used for the WTCHG WGS500 project (manuscript in preparation). The specifics of the Platypus variant detection algorithm have not been published. In brief, Platypus scans windows of the genome for variants that differ from the reference genome. Platypus detects variants with higher accuracy by using local realignment and assembly of sequencing reads to detect variants unlike other more established variant callers [100,107,108]. By realigning the reads in each window of the genome, Platypus does not rely on potentially suboptimal-alignments of reads that may produce false variants or genotypes [100]. This approach differs from many variant callers and has been shown to detect variants with greater accuracy. For the above reasons, Platypus was used as the variant calling algorithm in this project.

Platypus was run with a region file including the target exome regions with the command line input shown in Table 2.6. The output file format from Platypus was based on the established 1000 Genomes VCF 4.0 format. These files are tab-delimited and each begins with a series of meta-information lines describing the file format and the information contained in the specific INFO, FILTER and FORMAT fields followed by a header line. The header line contained the fields present in each line of data for the file. Platypus produces a line of data for each variant observed with the following fields: chromosome, position, reference allele, alternate allele, sequencing quality score (QUAL), filter, info, format and sample genotype information.

The QUAL is the Phred scaled probability that the call exists at that position based on -10*log(1-p) with a large number indicating a very small probability of an error [112]. The filter field contained either "PASS", indicating that the variant passed the specific filters applied by Platypus or it listed the filters that the variant failed, for example strandBias was applied when a variant was only observed in reads of one direction. The INFO field contained specific information and annotations for that position, such as total reads or population genotype frequencies. The FORMAT field defined what information was present in each sample genotype field. The default output sample genotype field contained the following fields GT:GL:GOF:GQ:NR:NV, where GT was the genotype 0/1 for heterozygous and 1/1 for homozygous, GL was the genotype likelihood metric, GOF goodness of fit value, GQ the conditional genotype quality, NR total number of reads and NV the total number of variant reads.

Table 2.6: Platypus variant caller command line settings
UNIX command line operations for variant calling with Platypus 0.5.2 in genotyping, variant calling, and trio variant calling mode.

Programme	Version	Command line
		Platypus in Genotyping Mode
Platypus	0.5.2	python Platypus.py callVariants --refFile=hs37d5 --bamFiles=*sample*.bam --source=SNP_ARRAY.vcf.gz --minPosterior=0 --getVariantsFromBAMs=0 -o *sample*.array.vcf
		Platypus in Variant calling Mode
Platypus	0.5.2	python Platypus.py callVariants --minFlank=0 --regions=/well/brc/brainmalf/Exome_v3_plus25.reg --refFile=hs37d5 --bamFiles=*sample*.bam -o *sample*.vcf
		Platypus in Variant calling for trios and families
Platypus	0.5.2	python Platypus.py callVariants --minFlank=0 --regions=/well/brc/brainmalf/Exome_v3_plus25.reg --refFile=hs37d5 --bamFiles=*father*.bam,*mother*.bam,*proband*.bam -o *family*.vcf

2.6.1. Family and trio based variant calling

To improve variant calling and *de novo* filtering, Platypus was run on all of the BAM files from family groups when available. In the case of individual samples without additional family members, Platypus was run on the single sample in variant calling mode. In the case of families either with suspected *de novo* inheritance or several affected family members present, Platypus was run as one instance on the whole family to generate a multi-sample VCF. This approach was used to find variants that might otherwise have been missed. When Platypus is run on several samples, it pools all the reads from all the samples for variant discovery (Figure 2.2). If a variant is in multiple samples then there is higher likelihood of detecting the variant. This is valuable for exome sequencing as coverage can be quite variable across the samples and exome. When a variant is discovered, genotyping is done on all samples and reported in the multi

sample VCF. For each variant reported there will be a genotype for each family member which allows for efficient filtering for *de novo* variants in downstream analyses.



Figure 2.2: Example of advantage of trio variant calling with Platypus. Shown is WES sequencing reads in IGV where there is read support for a variant in the proband (red T) but only one variant read in the mother. In this example, if Platypus were run separately on each sample the variant in the proband would have appeared *de novo* as it would not have been called in the mother with only a single read. When this family was called together with Platypus, all the reads were pooled for variant discovery and as a result the variant was called in the mother.

By calling the family together in one instance this also controlled for potential sequencing noise. A noisy locus might only get called as a variant in one individual if run separately, but by having Platypus call of the samples together the noise might be expected to be systematic and show up in more than one sample which could then be filtered out in IVA or by using genotype likelihood scores. Since it likely improved variant calling, Platypus was run as a single instance on all trios and families in this project.

2.6.2. Exome Region file creation

The method of targeting the exome was recommended for WES variant calling with Platypus and was the best practice established for genome sequencing analysis of WES by the Broad Institute for Genetic Research [113]. For the above reasons, only the exonic

regions targeted by the capture kits were used for variant calling in this project. To target only the exonic regions, a specific input file needed to be made for Platypus. The exonic target regions for the exome capture kits were downloaded from the respective manufacturers' website as BED files, a tab-delimited text file describing the position of each target. The SeqCap (NimbleGen) capture kit targeted all exonic regions of the 436 Genes of Interest (Appendix Table 1) and the TruSeq (Illumina) targeted 98.71% of all positions in those genes (Appendix Table 2). The target region BED file was modified to include 25bp upstream and downstream of each target. The modified region file and the original target region BED file were uploaded to UCSC as a custom track. A selection of regions was visually inspected to ensure that the file was modified correctly (Figure 2.3). The input region file for Platypus, which was simply a text file with chromosome:position-position, was created for each respective kit from the modified BED files.

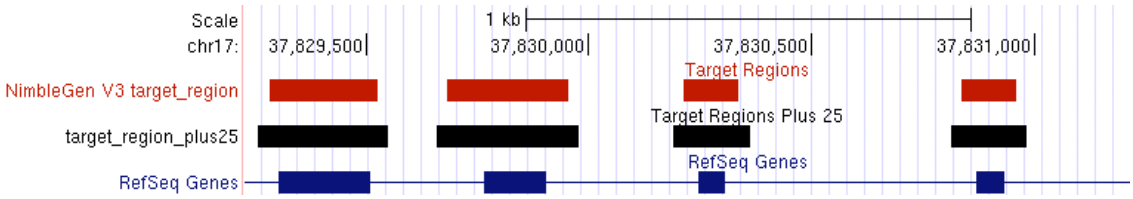


Figure 2.3: Example of modified exome target regions
The regions are shown as custom tracks in the UCSC Genome Browser. Target_region_plus25 (black) was used as the regions for variant calling with Platypus. It was based on the NimbleGen SeqCap EZ v3.0 target regions (red), including an additional 25bp at the start and end of each region.

The initial target regions specified by the capture kits did not contain any mitochondrial sequences. Since there were no specific probes designed for the Mitochondrial Genome (mtDNA) it was not part of the kit's target. This became apparent when an initial analysis

of the VCFs created with the region file found no mitochondrial variant calls. Although not directly targeted by the kit, the average mitochondrial coverage with WES is usually quite high. It was possible that the coverage was sequencing noise, which would increase the number of false variants in the data. For these reasons, the variant calling practice of the WTCHG Clinical Exome project does not include the mitochondrial regions. Many mitochondrial genes have previously been implicated in brain abnormalities[114]. Since there was high coverage and evidence suggesting that mtDNA calls from WES were real [115], mtDNA regions were included in the variant calling. The entire mitochondrial genome, MT:1-16569 (based on the revised Cambridge Reference Sequence) was added to the Platypus region file to allow for variant calling in the mtDNA regions.

2.7. Ingenuity Variant Analysis

A programme was required to efficiently analyse and filter the VCF data in this project. Ingenuity Variant Analysis (IVA) v3.0.2 was used in this project as it offered the ability to effectively filter data and also provided access to an in depth knowledge base. IVA is an online and cloud-based bioinformatics software, which is an advantage as variant annotation content and software are updated regularly and automatically. The efficacy and potential clinical application of IVA was assessed as part of this thesis.

To analyse the data, the VCF files from exome sequencing were run through a pipeline on the WTCHG severs and uploaded directly to IVA by the UNIX command line FTP programme curl. The transfer of data from the secure access servers to the cloud storage was encrypted with the Secure Socket Layers protocol to ensure the data was protected.

2.7.1. Modification of VCF files for Ingenuity Variant Analysis

The genotype fields from Platypus are not the standard VCF format and to import the VCFs into IVA they needed to be modified. For these reasons a custom python script was written to add the necessary genotype flags. The genotype field from Platypus 0.5.2 is GT:GL:GOF:GQ:NR:NV. The NR and NV fields are not defined in the standard VCF 4.0 formatting by the 1000 Genomes. These fields were developed as part of the Platypus calling algorithm and were useful for direct command line variant filtering. In VCF 4.0, there was no standard field for variant read depth, NV, and the standard flag for depth was DP, not NR. Since the primary tool used for variant filtering was IVA, which only recognizes the standard genotype flags, the genotype fields needed to be modified so that read depth and allele depth were available for the subsequent analyses.

A Python script NR_to_DP_AD_VF.py was developed as part of this project to add the necessary genotype field flags to each VCF. The programme was based on a Perl script vcf_format_AD_DP.pl developed by BRC bioinformatician Dr Gareth Wilson. The Python script was developed to deal with multi-sample VCFs and multi-allelic variants. In simple terms, the Python script loaded the input VCF file and read the header to determine the number of samples present. A new file was created with the original header and new meta-information lines for the Allele Depth (AD), the Total Depth (DP) and the Variant Frequency (VF) which explained each value and how it was calculated.

After the header was modified, the programme read through the VCF line by line. When the programme read a line it split the VCF fields into variables for each sample genotype

present. The function which modified the sample genotype split it into the relevant fields: GT:GL:GOF:GQ:NR:NV. The function calculated the AD as the NR value subtracted by the NV value and the NV value. The DP at the position was reported as the NR value. The variant frequency VF was the NV value divided by the total number of reads (NR). In the case of multi allelic variants the calculation was more complex, as multiple variant read counts were used to calculate the AD, VF and DP. Once the fields were calculated, the genotype fields were concatenated back into a single variable with the new AD:DP:VF flags on the end of the existing genotype. The programme concatenated the variables and the sample genotypes into a single line, which was outputted to the new VCF file. The Python script was validated by comparing its output against the Perl programme and an unmodified VCF.

2.7.2. Ingenuity VCF import and annotation pipeline

An important step in variant identification is the annotation of variants. The specifics of the IVA proprietary pipeline have not been published. Based on experience with the software and personal communication with Ingenuity, the VCF data is imported by the following method. The pipeline first reads the header of the VCF file to determine information stored in the INFO fields. Each line of the VCF is read for variant position, reference and alternate allele. Any annotation present in the ID column or non-integer information in the INFO field is discarded. Sample specific information including genotype, read depth, and allele depth are imported from each sample genotype field. Sequencing quality and filter status for each variant are also imported. Homozygous reference (0/0) genotypes are not imported and no call variants (./.) are imported but only reported in analyses alongside samples with variant calls at the same position with the

flag “no genotype”. Multi-allelic variants are split into two variants differing from the reference at the same position. In the case of multi-sample VCFs, IVA imports the VCF and separates it into individual files before annotating each position.

Variants are annotated in IVA based on position, reference and alternate allele, with the following content sources: Ingenuity Knowledge Base (Arrakis 140408.002), COSMIC (v68), dbSNP (Build 138 (08/09/2013)), 1000 Genome Frequency (v3), TargetScan (v6.2), EVS (ESP6500 0.0.21), JASPAR (10/12/2009), PhyloP hg18 (11/2009), PhyloP hg19 (01/2009), Vista Enhancer hg19 (12/26/2010), CGI Genomes (11/2011), SIFT (01/2013), BSIFT (01/2013), TCGA (09/05/2013), PolyPhen-2 (HumVar Training set 2011_12), Clinvar (02/11/2014).

2.7.3. Default Ingenuity Variant Analysis filtering

The principal purpose of utilizing the IVA software was to be able to access the knowledge base to filter for rare variants in genes previously implicated or associated with brain abnormalities based on the biological context of the disease. Once the samples were uploaded and imported to IVA they were presented in the online user interface for analyses. Proband and family members were analysed together. Pedigree files were uploaded and associated with the relevant families to allow for filtering based on inheritance. The genetic disease option for the analysis was used in all cases. The specific inheritance pattern selected for the initial analysis depended on the case and family history. In cases of suspected *de novo* inheritance, a dominant model excluding variants present in the parents was selected. A recessive model was used for any cases with suspected consanguinity or family history not suggestive of dominant inheritance.

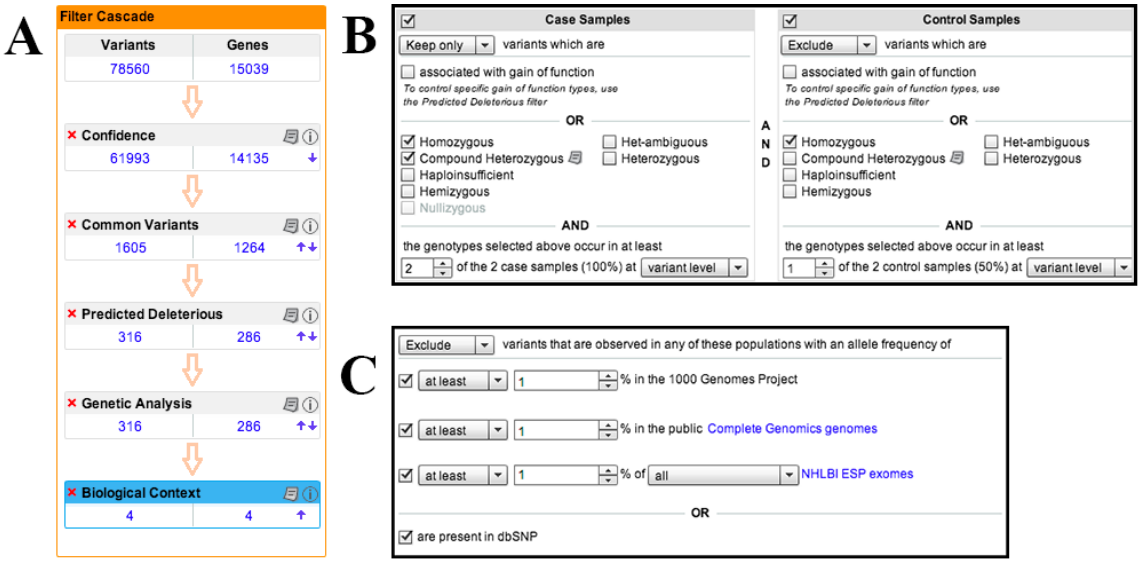


Figure 2.4: Filtering cascade and filtering options in IVA
(A) IVA filtering cascade showing numbers of variants at each stage of filtering. The 5 main filters are shown sequentially in the cascade.
(B) Genetic Analysis filtering showing options for different inheritance models. Different options can be selected based on the suspected inheritance.
(C) Common Variant filtering with options to exclude variants at frequencies in common variation databases and those reported in dbSNP.

The initial analysis contained five main filters: Confidence, Common Variants, Predicted Deleterious, Genetic Analysis and Biological Context. The user interface in IVA presented the filters in a cascade (Figure 2.4) with the total numbers of variants present at each filtering step, which could be queried and viewed separately. The Confidence filter selected variants of high quality. The initial setting was to filter based on a minimum call quality of 20 (1 in 20 chance of error). The Common Variants filter excluded variants observed at specified frequencies in v3 of the 1000 Genomes Project, Complete Genomics Genomes [21] and the ESP6500 version of NHLBI ESP exomes [22]. The Predicted Deleterious filter selected variants with either predicted or observed evidence suggesting they could disrupt gene function or expression [116]. There were two main

aspects to the filter, selecting variants observed to be associated with a phenotype and those observed or predicted to cause a net gain-of-function or net loss-of-function in a gene.

The Genetic Analysis filter selected for variants based on the specified inheritance model from the initial analysis setup. In the case of a simple recessive model of inheritance, the filter would select variants that were either homozygous or compound heterozygous in the proband and not homozygous or compound heterozygous in the parents or controls. The default filtering settings also included haploinsufficient, hemizygous and het-ambiguous variants. In the case of dominant or suspected *de novo* inheritance the filtering selected for heterozygous variants present in the proband that were not present in any form in the parents or control.

The final filter was for the Biological Context, which selected for variants in genes associated with the key term provided in the initial analysis setup. The biological context filter accessed the Ingenuity Knowledge Base, which has been continuously curated and updated since its first iteration as a database for Ingenuity Pathway Analysis in 2005 [117]. The knowledge base was formed by a process of content acquisition from high-quality peer reviewed scientific journals and medical databases [118]. In all cases the information was manually reviewed by a group of experts at Ingenuity to ensure that the content was accurate [118]. There were several levels to the biological context filtering; in its simplest form it was a list of genes associated with key phenotypic terms. Beyond the direct associations, the database was a network of key terms and biological evidence

including animal models, epigenetic studies, information on clinical trials and specific tissue context and processes. The filter included the option to select for variants directly associated or within one or more biological steps of the key term. The biological steps could be simple, ie if gene A was associated with the term and there was evidence that it binds gene B, gene B would be listed within 1 step of the key term (Figure 2.5). There were also more complex associations based on known gene and biochemical pathways, including both upstream and downstream regulators. The default settings for the Biological Context filter selected for variants directly associated with the provided key term. If there were no variants found, the default would select for variants within 1 or 2 biological steps.

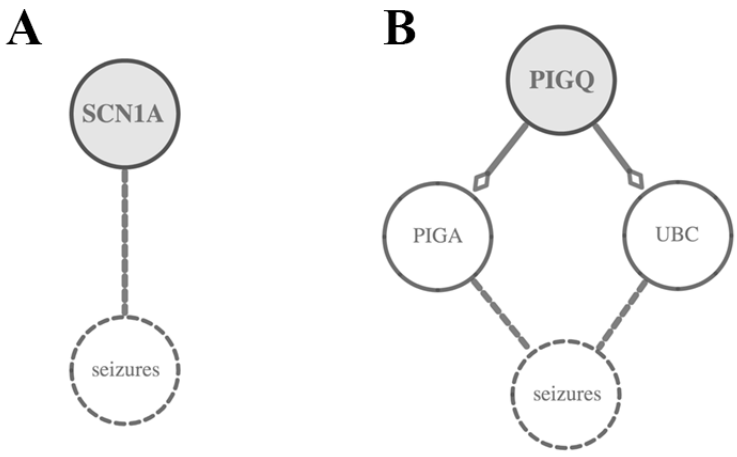


Figure 2.5 Examples of biological context filtering in IVA
Shown is the graphical representation in IVA of two biological context relationships. The key term, seizures, is shown at the bottom, with connections between genes and the key term represented by lines. Each line can be queried to view the association.
(A) 1 step relationship in IVA where *SCN1A* is directly associated with seizures since it is a Dravet Syndrome disease gene [119].
(B) 2 step relationship in IVA where *PIGQ* is associated with seizures based on it being in the same pathway as *PIGA* (GPI Pathway) and *PIGA* causing a GPI disorder with seizures [120]. Association with UBC is based on protein binding.

2.7.4. Validation of Ingenuity Variant Analysis

Since IVA was not part of the standard WTCHG bioinformatics pipeline, the efficacy of the software was assessed. The strategy to validate the software was to analyse existing data. This included data from several WGS500 projects, both published [94] and without any strong candidates. The annotated VCFs from freeze5 of the WGS500 projects were uploaded to IVA utilizing the same pipeline as described above. The data from the WGS500 projects was analysed without any knowledge of candidate genes previously identified with UNIX based bioinformatics tools and scripts. The efficacy of IVA was assessed based on the identification of the candidate variants and overall accuracy of variant reporting. Accuracy of the variant reporting was evaluated by examining a selection of raw Platypus calls against the variants as reported in IVA.

2.7.5. Custom IVA filtering developed for variants associated with brain abnormality phenotypes

A strategy was developed to facilitate filtering in IVA and efficiently identify candidate variants. Since the generic filtering in IVA is very broad, it was modified to meet specific requirements of the brain abnormality project. The standard filtering was also adapted to take full advantage of the features present in IVA.

To control for sequencing noise, custom annotations were imported from the INFO fields of the VCF, which included the total number of number of forward and reverse variant reads. A custom filter was used to exclude variants that were only observed in one direction in the sequencing. WGS500 allele frequencies were also imported as an

additional population frequency to filter on in the downstream analysis. To exclude genes known to show high degree of exonic variation, which contributes to sequencing noise, the filtering settings were modified to exclude variants in the top 0.2% of most exonically variable 100base windows in healthy public genomes and outside the top 1% of most exonically variable genes from the 1000 Genomes project. In order to control for common variation, the default variant allele frequency filtering was modified from 3% to 1% exclusion frequency for the initial analysis. This was in part based on the prevalence of the main inclusion criteria, microcephaly, which can range from 1/1,000,000 (estimates of frequency in Caucasian populations) to 1/10,000 in those originating from Northern Pakistan [74]. A 1% allele frequency was also chosen since it was a balance between the rarity of the phenotypes and the possibility of common variation contributing to the phenotypes. Since there was no clinical associated dbSNP list available for filtering in IVA, all variants present in dbSNP were also excluded to control for common variation. These settings were an effective method filter out common variation and noisy variant calls in the data for the first pass analysis.

The initial analysis focused on variants that were predicted to cause a net loss-of-function. For this reason, the deleterious filter was modified to only include those variants predicted to alter protein sequence (missense, nonsense, stop loss, indels, and splice site variants within 2bp of intron exon border). The data was filtered based on the presumed inheritance pattern in the family with the Genetic Analysis filter. In the case of autosomal recessive inheritance, the settings were modified to not filter out variants that were potentially compound heterozygous in the parents or controls as this could erroneously

eliminate compound heterozygous candidate variants in the proband. An example of this this occurred when a patient had compound heterozygous variants, one inherited from each parent, but the variants were excluded by the Genetic Analysis since one of the parents was also compound heterozygous in that gene but for a separate common missense change.

The Biological Context filtering was set to "directly associated" with a single key term. In most cases the initial key term was "brain abnormalities" which included a list of 436 genes previously associated with brain abnormalities (Appendix Table 1). Terms were also selected for filtering based on the patient's specific phenotype. Based on this filtering strategy, an initial set of candidate variants could be quickly determined and assessed.

If no strong candidates were present in the initial analysis, the filtering was broadened. The first step was to include conserved regions (phyloP p-value less than 10^{-7}), potential promoter sites in the UTR and variants in Encode consortium defined transcription factor binding sites [121]. The second step was to expand the Biological context filtering to include further key phenotypic terms and relationships.

To be comprehensive, the data was filtered in all cases disregarding inheritance for any rare deleterious variants in known genes for microcephaly and brain abnormalities. This was done by directly filtering all of the variants predicted as deleterious with the Biological Context filter. The list of variants was carefully scrutinized for any strong

candidates that were missed after the Genetic Analysis inheritance filtering. The filtering was also modified to include dbSNP variants predicted to cause protein changes. These were examined to ensure there were no dbSNP candidates reported as causative for brain abnormalities present in the sequenced individuals, ie rs145489194, a stop gain variant in *ASPM* reported in the NCBI Database of Clinical Variation (ClinVar) [122] and OMIM and pathogenic. As a final step to ensure that no potential variants were missed, the Predicted Deleterious filter was removed from the cascade. The list of variants selected by the Biological Context and Genetic Analysis filters was examined for any potential candidates that were mis-classified as non-deleterious.

Variants were classified as candidates based on the predicted effect on protein coding and the biological evidence associating the gene with the disease. All of the original source article evidence for the association of a variant with the phenotype from the IVA Biological Context filtering was examined before a variant was classified as a candidate. The frequency of the variant was examined in the WGS500 allele frequencies and the EVS Exome Variant Server. Variants that were previously reported required stronger biological evidence to be classified as a candidate.

2.7.6. Incidental findings

In this project, the exome sequencing data was not screened for incidental findings of clinical relevance. This is contrary to the American College of Medical Genetics (ACMG) and Genomics recommendations that a set of genes should always be screened for pathogenic variants and reported back to patients [123]. Recently the ACMG updated the policy stating that patients should be able to decide if they wish to know of incidental

genetic findings [124]. Under the current ethical framework of the Oxford Brain Abnormality Research group, there is no provision to deliberately screen the data for incidental findings. If a truly incidental finding is identified inadvertently, the families are only informed of such incidental findings if they requested to be informed at the time of their recruitment. However, the chance of such incidental findings here is low in view of the IVA filtering. There is considerable on going debate about NGS and incidental findings [125], therefore it is likely that this policy will be updated to match the current trends when the ethical approval for the research group is renewed in the future.

2.8. Candidate variant examination in IGV

When a candidate variant was found with IVA, the quality of the variant was examined to ascertain if the variant was real or sequencing noise. To assess the variant, the aligned sequence data (BAM files) containing candidate variant was examined using the Integrative Genomics Viewer (IGV) [126]. The coordinates and immediate region surrounding each variant was examined visually. The visual examination was to ensure that candidate variants were not potential sequencing artifacts (Figure 2.6) or variants incorrectly called or mis-annotated by Platypus. Variants in regions with low mapping quality or containing a significant amount of mis-mapped reads were carefully scrutinized. The regions in question were also examined in UCSC Genome Browser for the presence of segmental duplications or other repetitive elements which could cause sequencing reads to be incorrectly mapped. The Genome Browser was also used to confirm that the variants were not common variation reported in dbSNP and other database but mis-annotated by IVA. The high quality variants that appeared to be real in

IGV and not common variation were put forward as candidate variants to be validated with Sanger sequencing.

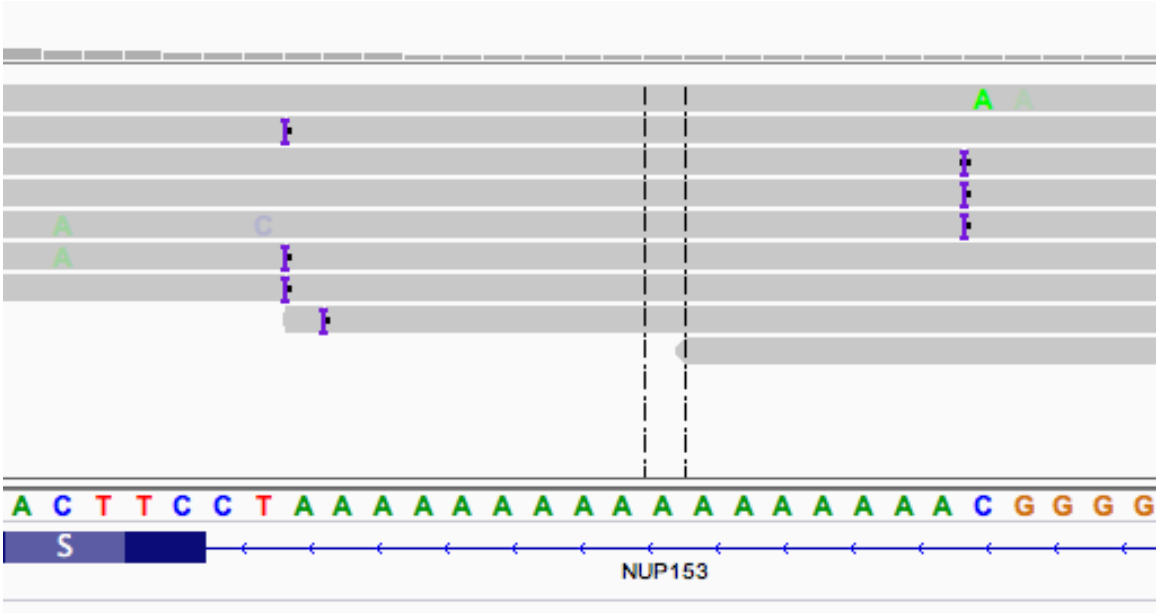


Figure 2.6 Likely sequencing artefacts due to poly A nucleotide repeat
Shown is a probable sequencing artefact in IGV. The grey bars represent WES sequencing reads and the purple I's are insertion of A. The sequencing noise is due to the homopolymer repeat which is poorly captured by NGS [127].

2.9. PCR primer design for candidate variants

In this project, all candidate variants identified were validated with Sanger sequencing. To validate a candidate variant, the genomic region surrounding the variant was amplified with the Polymerase Chain Reaction (PCR) [19,20]. To effectively design primers in a reproducible manner, the following method was used; a BED file was created with the candidate variant coordinates and imported into UCSC Genome Browser [128] as a custom track. Alongside the custom track, certain specific tracks curated by UCSC for use in the genome browser were selected for primer design: RefSeq Genes track based on NCBI Reference Sequence Database Release 64 [129], the NCBI database of genetic

variation dbSNP (138) including all known SNPs [130], Segmental Duplication [131], Repeating Elements by Repeat Masker [132] and Tandem Repeat Finder [133] (Figure 2.7). These tracks were selected to prevent primers being placed in repetitive sequences, or designing primers that might not match the subject's target DNA sequence due to common variants. The browser settings were saved as a specific UCSC session to ensure that primers selection and design was consistent between variants.

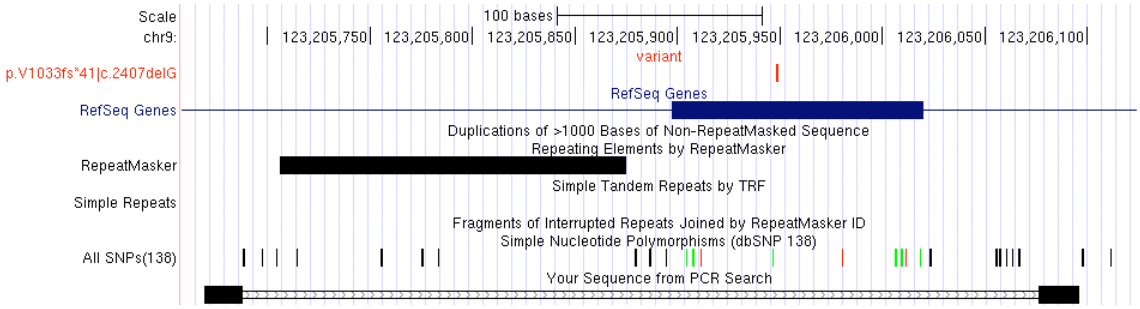


Figure 2.7: Primer design in UCSC Genome Browser.
Primer design for *CDK5RAP2* candidate variant is shown in UCSC Genome Browser. The black bars in bottom track represent the forward and reverse primer sequences.

The extended DNA sequence of the target region was obtained with specified colours indicating the variant and the repetitive sequence tracks. The target sequence was submitted to the online tool Primer3 version 4.0 [134,135] with the standard settings, avoiding sequences based on the human mispriming library [136]. The target region was defined as being 80bp upstream and downstream of the variant to allow for efficient Sanger sequencing of the region. SNPs or other repetitive elements present were marked for Primer 3 as regions to exclude from potential primers. The primer sequences obtained were run through the Genome Browser In-Silico PCR tool and the predicted product was compared against the variant and expected product from Primer3. The specific primer

sequences were also examined in the Genome Browser as a final check to confirm that they did not contain any common variants or were within repetitive elements. In the case when primer placement could not be reliably obtained outside of repetitive elements, one primer was placed in the repetitive element and one outside. If this approach was not possible a long-range nested PCR strategy was undertaken to sequence the variant. Once the primer design was validated In-Silico, they were ordered as unmodified oligonucleotides purified with the High Purity Salt Free (HPSF) method from Eurofins Genomics, Ebersberg Germany (Table 2.7).

Table 2.7: PCR and Sanger sequencing primers for candidate variants

Patient	Gene	Primer	Reverse	Product Length
BRC_115	PCDH19	CCTGAGTCGCGATCAGACA	CTGTCACTGGCGCTTTAGAC	241bp
BRC_16, BRC_59	CASK	GCCCACACTGCTAAAGACAG	ACATTGATGCCTCTGGATTTTGT	380bp
BRC_81	CD5K5RAP2 - c.4604+1G>C	GTTTGTGTCTACTGCCTGCC	GCCAGCGTGAAGGAAGAAAA	233bp
BRC_81	CD5K5RAP2 - c.3097delG	ACGGGTGGGTTCTGACATC	TTCTTCTCAGTTGCTTGGCT	427bp
BRC_90	ASPM - c.10086delA	AAAAGCAGGTTTGAACACACA	ACTTGCTGCAGGTGGTCC	432bp
BRC_90	ASPM - insertion	TCGAACAATTAACCGCCTAGC	AAGGAGTAAGCTTGGAGGGG	349bp
BRC_41	PGAP3	TTACCCCATGCCTCAGACTG	TTCACCAGGCCACATTTGTC	227bp
BRC_146	SOX11	CCTGCAGCGGATCATGGT	CCATGATCTTCCTGCGTTCG	218bp
BRC_224	PSMB1	GTGACTCTCTAGCTATCTGACCA	GAATACAGACCCCTCCTCCC	337bp
BRC_225	CHMP7	GCTACCCTCCTGACTTCCG	CCCATCAACTCTGTCAGCCA	397bp
BRC_230	GPT2	CCTGGGAGCTGTCATCCTC	CCTGGGAGCTGTCATCCTC	295bp
BRC_226	CLIP2	GGGGCTAATGGACAACTGGA	GGCTTGCACCTCCAGCTG	313bp
BRC_91	PHF7	TGAAAACTCAGGGGACATCC	GGGACAGCAGCTACTCAGAA	241bp
BRC_102	PIGY	CCAAGGCCAGCGAGTTAC	CCGGTGAAAACCCAGACAC	526bp
BRC_179	CLASP1	GATGGCATCTGAGTAGGGGT	TGTCACTGGCCATACATATTTCA	420bp
BRC_50	ACADSB	GCCCATGTTGCAATTCTGAG	CTCCTGTGGTTCCATGCAAG	480bp
PMG fam	PI4KA - Missense	AGAAGCCCTAATTTACCCCGT	CTAACAGCGGCCTCTCTCC	290bp
PMG fam	PI4KA - Missense V2	CCGTGGCACCTGAACCATA	same as above	273bp
PMG fam	PI4KA - Stop Gain	GTGGGCTCTGACCTCACC	ACCCAGTCTGAGTTTCTGAGA	300bp
PMG fam	PI4KA - LR A	CGTGTGAAGTAGTCGTACATGC	CATTTATCTAGATGAAGAGGGCCA C	8810bp
PMG fam	PI4KA - LR B	CTCCCACACTCAGAACTTAACTTTC	ATTGGTTTGAGAGTTGTGCTTCTTA	10047bp

2.9.1. PCR amplification of candidate variant regions

Candidate variant regions were amplified by PCR from patient DNA samples using the following protocol. The FastStart Taq DNA Polymerase system from Roche was used for the PCR reaction which was developed based on the "Hot Start" method with thermally activated DNA polymerase enzymes [137]. Standard PCR reaction mix consisted of 2.0µl of genomic DNA (~20ng/µl), 2.5µl of 10X PCR Buffer, 1.0µl Primers, 0.5µl PCR nucleotide mix, 0.25µl FastStart Taq DNA Polymerase and 18.75µl of distilled membrane-filtered H_2O from Gibco. Final concentrations in the 25µl PCR mix were 2mM $MgCl_2$, 200µM of each dNTP, 0.4µM of forward and reverse primer and 1.25 units of polymerase. The PCR reactions were run on a peqSTAR 96X Universal Thermal Cycler from peqlab (Erlangen, Germany). Standard PCR temperature profile was as follows, heat thermal cycler lid to 110°C to seal PCR plate film, 95°C for 2 minutes to activate the polymerase and denature the DNA, followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing at 57°C for 30 seconds, elongation at 72°C for 30 seconds and a final elongation step at 72°C for 6 minutes. The size of the PCR fragments were confirmed with a 1.8% agarose gel containing 0.25µg/ml of ethidium bromide which was run at 100V for 30 to 60 minutes before being visualized with a BioRad ChemiDoc MP Imaging System under ultra violet illumination. HyperLadder 100bp (IV) was used as the size standard.

For certain reactions the PCR conditions were optimized by varying the annealing temperature from 54°C to 60°C based on the suggested primer annealing temperatures from Eurofins. For regions with greater than ~65% GC content, 5.0µl of GC-Rich solution was used in place of 5.0µl of H_2O which works by adding DMSO to the reaction

mix and increased the final concentration of $MgCl_2$. The DMSO and increased $MgCl_2$ concentration is thought to alter the melting behavior of the DNA to allow for efficient priming and PCR in GC rich regions. In most cases the GC rich solution was used after the first set of PCR reactions failed and high GC content was identified with the Genome Brower GC percent track. For all reactions and primers designs, a template negative control was also run alongside the reactions to control for contamination and potential primer dimerization.

2.10. Sanger sequencing of candidate variant regions

In this project, I validated all of the candidate variants with Sanger-sequencing. Although WES is a well-established method, it is still standard practice to validate all variants by traditional sequencing. After the PCR reactions, the primers and remaining dNTPs were removed with the EXOSAP method [138]. The EXOSAP reactants as follows were added to each PCR reaction; 0.1µl of Exonuclease (Exo) I enzyme from New England Biolabs, 1.0µl of Shrimp Alkaline Phosphatase (SAP) from USB Affymetrix, 1.0µl of 10X SAP buffer and 0.9µl of distilled H_2O. Thermal conditions for the reactions was as follows, thermal cycler lid heated to 110°C, 40 min incubation at 37.0°C to digest single stranded DNA (primers) with Exo I and to dephosphorylate any unincorporated dNTPs with SAP, followed by 15 min at 80°C to deactivate the enzymes.

Sanger sequencing [18] (di-deoxy) reactions were done with the BigDye Ready Reaction Mix (Applied Biosystems). The sequencing reactions were prepared with 1.0µl of the PCR reaction post EXOSAP, 0.5µl of 3.3µM forward or reverse primer, 1.0µl of the

Ready Reaction Mix, 1.5µl of 5X Sequencing Buffer and 6.0µl of distilled H_2O. Reactions were run in standard 96 well PCR plates with the peqSTAR thermal cycler. The thermal conditions were as follows, lid heated to 110°C, 96°C for 1 minute to denature the DNA, followed by 35 cycles of denaturation at 96°C for 30 seconds, annealing at 50°C for 15 seconds, and elongation/termination at 60°C for 4 minutes. Following the reaction, the samples were ethanol precipitated in preparation for sequencing. The following was added to each reaction, 2µl 125mM EDTA, 2µl 3M NaAc and 50µl of 100% ethanol. The mixture was left at room temperature for 15 minutes before being centrifuged at 3000xg and 4°C for 30 minutes. The plate was inverted to remove the supernatent and then and spun upside-down at 185xg for 10 seconds to remove any remaining liquid. A wash step was then performed with 70µl of 70% ethanol added to each reaction before being spun at 1650xg and 4°C for 15 minutes. The final mixture was inverted and spun at 185xg for 10 seconds before being left at room temperature for ~ 1 hour to dry.

Precipitated reactions were sent to the Oxford Zoology Core Sanger sequencing facility to be sequenced by Capillary Electrophoresis on the Applied Biosystems 3730xl DNA Analyzer or 3130 Genetic Analyzer. At the Core facility, the reactions were re-suspended in formamide and centrifuged before being injection into the capillary of the sequencer [139]. The sequencing was done based on the principal of dye terminator chemistry. Sequencing reactions resulted in fragments of differing lengths due to the presence of normal dNTPs and each of the four di-deoxy nucleotide terminators tagged with a different fluorescent dye [140]. The products of each reaction were run on a single

capillary and a high voltage charge was applied to separate the extension products based on size. When each extension product passed the positive electrode, a laser caused the dye to fluoresce and the resultant signal was recorded and converted by the Genetic Analyzer software into digital chromatogram files (ab1 file) [140].

2.10.1. Analysis of Sanger sequencing data

The data from the Sanger sequencing were downloaded directly from Zoology file servers using the open source programme CyberDuck v4.2.1. The chromatograms were analysed visually with Chromas [141] and the 4Peaks software package [142]. Fasta formatted text files were made containing the exported sequence from 4Peaks for each individual or family sequenced for the specific candidate variant. The fasta file for each variant was uploaded and run through the UCSC Blat Tool [95]. The blatted sequences were compared against the Refseq genes and the custom track based on the candidate variant bed file used for the original primer design. UCSC Blat tool highlighted those positions that were non-concordant with the reference sequence. If the blatted sequence matched the targeted region with a high degree of concordance, the chromatogram traces were inspected for the expected variant. In those cases with suspected *de novo* inheritance, the parental traces were examined at the variant position to ensure that there were no potential second peaks suggesting heterozygosity or mosaicism in one of the parents. The variants that appeared real in the chromatogram traces and matched the predicted segregation from the exome sequencing were considered validated.

2.11. Sequencing of *PGAP3*

In this project *PGAP3* was identified as a potentially novel disease gene. Sanger sequencing primers were designed to sequence the whole gene to ascertain if other patients with similar phenotypes harbored deleterious variants in *PGAP3*. The entire coding sequence of *PGAP3* was amplified from genomic DNA and sequenced with standard conditions as described previously. Primers listed below were designed as described above, except that the target regions were defined as 80bp upstream and downstream of the coding region of all eight exons of the gene (Table 2.8). Conserved regions within the 5' and 3' UTRs were also targeted. Conditions for the sequencing were modified to include the GC rich additive for exon 1 and 8 due to the high GC content. The chromatograms from the sequencing were analysed visually with the 4Peaks software [142]. Consensus sequence was compared to the Refseq *PGAP3* transcript variant 1 (NM_033419.4) using the UCSC Blat Tool [95]. The Linux based tool Consed [143] was used to analyse the data across multiple samples which has a built in package PolyPhred [144] which can reliably detect heterozygous SNPs. All of the variants identified were compared against known variants reported in dbSNP, 1000 genomes and the WGS500.

Table 2.8: Forward and reverse *PGAP3* PCR and Sanger sequencing primers

Exon	Forward	Reverse	Product Length
1	CCTCGTGGGGAAACTAAGGT	GACTTCAGCGGAGAGTAGAATAA	391bp
2	TCCAGAGGCAGGACACATG	ACACCTTCCCCTATGTGTACA	477bp
3	GGGAAACAGAGGGGAAGCT	CCAAACTCCAGAGCTGAGGA	386bp
4	ATGACCCTCTAGAGCCCCTC	GTTTGTTCAGAGGGCTCCAG	295bp
5	CTCACCCAGGACAGCTGAT	CCTTGTTCCTGGGCACATTT	392bp
6	TCCTCTCTCTCCCACCTCAA	GAGGACTCTGGGGTAGGACT	483bp
7	CTCTGACAGCTCGCATGG	TGGGGTAGAGGAATGGAGAAG	510bp
8	GGTAGACAGAGGGCTTCAGG	TTCCTTGCCCACTCCTTAGG	574bp

Chapter 3: Results

In the Results there are two main sections, the development of the bioinformatics pipeline (section 3.1) and the application of that pipeline for the identification of variants associated with brain abnormality phenotypes (section 3.2). The pipeline results cover the use of several bioinformatics tools and the overall quality assessment of the exome sequencing data. The candidate variant section focuses on the results of using the pipeline and the variants of interest identified.

3.1. Bioinformatics and variant calling pipeline

A pipeline was developed in this project for analysing exome-sequencing data of brain abnormality patients. This pipeline built on existing elements of the WTCHG Core bioinformatics pipeline. Ingenuity Variant Analysis (IVA) was the main VCF annotating and filtering tool used in this project. Its efficiency and overall efficacy were assessed as part of this thesis. The pipeline was designed in part to address the unmet need for a robust pipeline that produces data that can be analyzed without using traditional command line based tools.

3.1.1. Comparison of autozygous region mapping with PLINK and HomozygosityMapper

Since the HomozygosityMapper (HM) is a relatively new tool, its output was compared with the established command line tool PLINK. The autozygous regions were mapped from SNP array data in the four consanguineous samples as described in the Methods section 2.4.2. The homozygous regions were concordant between PLINK and HM as all 9 autozygous regions were detected by both programs (Table 3.1). The only major difference between the tools was how the centromere of chromosomes was interpreted

(Figure 3.1). Another difference was that the exact estimates of the start and end positions of the regions differed between PLINK and HM by a small amount. HM would usually estimate the region as slightly larger than PLINK. The most noticeable example of this was for the chromosome 17 region detected in BRC41/42/216, which was 875kb larger when estimated with HM. To not introduce a source of potential error, all regions were inspected visually in Nexus before deciding the exact coordinates to use for downstream analysis.

Table 3.1: Concordance of shared homozygous regions between HomozygosityMapper and PLINK
Shown are the shared homozygous regions in affected siblings calculated by HM and PLINK in the three consanguineous families run on SNP arrays (Table 2.1). Chromosome 2 was called as two events with HM in BRC_179 due to rs13000417 being called as heterozygous in the array data.

		HomozygosityMapper			PLINK		
Patient sample	**chr**	**start**	**end**	**length**	**start**	**end**	**length**
BRC41, BRC42, BRC216	17	31,402,338	45,184,794	13,782,456	31,385,743	44,293,020	12,907,277
BRC102,103	4	88,485,747	96,213,533	7,727,786	88,502,642	96,206,860	7,704,218
	10	0	7,713,981	7,713,981	135,708	7,718,053	7,582,345
	12	21,505,093	23,091,547	1,586,454	21,512,197	23,079,828	1,567,631
BRC179,BRC180	2	101,192,499	110,086,000	8,893,501	101,200,213	127,359,823	26,159,610
	2	110,086,000	127,376,895	17,290,895			
	8	97,664,572	118,300,497	20,635,925	97,761,434	118,289,276	20,527,842
	11	37,080,974	46,174,945	9,093,971	37,095,080	51,530,241	14,435,161
	11	46,174,945	61,655,305	15,480,360	54,835,623	61,627,881	6,792,258
	12	8,704,521	11,761,718	3,057,197	8,718,136	11,754,985	3,036,849
	17	27,744,521	29,744,499	1,999,978	27,754,981	29,757,249	2,002,268

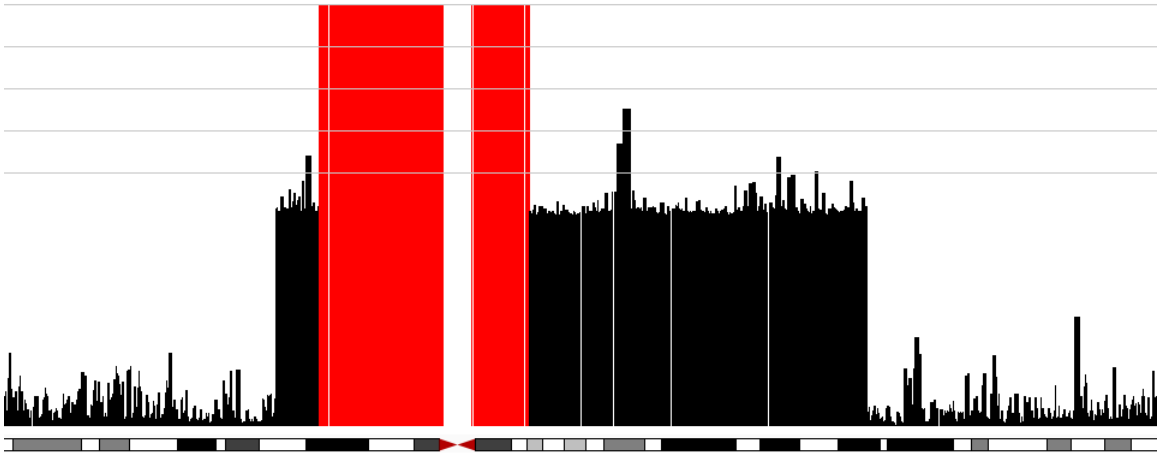


Figure 3.1: Shared homozygous chr11 regions of BRC_179 and BRC_180 from HomozygosityMapper
Shown is the output from HM where the chr11 homozygous regions were interpreted differently from PLINK. Red represents the shared ROH in BRC_179 and BRC_180. Chromosome bands were added to the output from HM to show the position of the centromere (the gap between the two large red regions). PLINK also detected this as two regions but included less of the centromere in the autozygous region. In practice, regions such as this were considered to be a single contiguous region for downstream analysis.

3.1.2. Comparison of estimation of Coefficient of Inbreeding with reported degree of consanguinity

The coefficient of inbreeding was estimated for the reported consanguineous individual run on SNP arrays. This calculation was based on the total sum of homozygous regions divided by the total length of the autosome in individual as described in the Methods section 2.4.3. The calculation functioned as an approximation of the reported degree of consanguinity for the families (Table 3.2). In the case of BRC_41, the coefficients of inbreeding highlighted the high degree of consanguinity in the father and that the parents were not first cousins as initially reported by the referring clinician.

Table 3.2: Comparison of estimation of coefficient of inbreeding with reported degree of consanguinity
Shown are the coefficients of inbreeding for the consanguineous samples run on SNP arrays.

Patient ID	Parental relationship	Expected coefficient of inbreeding from pedigree	Total sum of homozygous regions > 4Mb	Estimated coefficient of inbreeding
BRC_179	First cousins	1/16	223,929,782	1/12.7
BRC_180	First cousins	1/16	146,313,918	1/19.4
BRC_50	First cousins	1/16	154,155,028	1/18.4
BRC_41	Second cousins	1/64	60,355,989	1/47
BRC_42	Second cousins	1/64	61,833,579	1/46.9
BRC_216	Second cousins	1/64	50,796,290	1/55.9
Father(BRC_41)	unknown		224,736,452	1/12.6

3.1.3. Effect of the reference genome on variant calling

The variant calls of samples mapped to hs37d5 and Human37 were compared to ascertain if hs37d5 should be part of the pipeline. The first cohort was remapped to hs37d5 as described in the Methods section 2.5.1. The total number of variants called in the samples decreased when mapped to the hs37d5 reference compared with Human37 (Table 3.3). A small portion of the variants uniquely called in the Human37 reference VCFs passed the filtering by Platypus (QUAL > 200). The assumption was that these variants were due to mis-mapped reads, which would have been mapped to the decoy chromosomes in hs37d5. To assess the validity of the assumption and the quality of these calls, the unique X chromosome calls in Human37 in the 5 male samples, which were mapped to both genomes were examined (Table 3.3).

Table 3.3: Comparison of total number of variants called in samples mapped to Human37 or hs37d5.
Shown are the number of variants called in the 10 samples mapped to both genomes and the number of variants unique to Human37. The numbers of unique high quality variants on chromosome X and the zygosity are listed for the male samples.

Human 37	hs37d5	Number of high quality variants uniquely called in h37	chrX variants unique to h37	chrX variants unique to h37 and called as heterozygous
72954	69640	1558	41	41
88517	84078	1362	39	39
69027	66349	1424	70	68
76435	73298	1424	47	47
91071	86211	1421	58	55
78113	74804	1431		
74624	71513	1465		
89436	84772	1343		
85193	80956	1343		
89312	84344	1476		

There were 255 variant calls in the 5 samples on the X chromosome, of which 98% were heterozygous. Since the samples were male, the heterozygous calls were likely due to mis-mapped sequencing reads (Figure 3.2). The few unique homozygous calls were common to several unrelated samples (Figure 3.2), suggesting that they as well were due to mis-mapped reads. Based on the analysis of the X chromosome calls in the 5 male the samples, and the total decrease in variant calls in all 10 samples, mapping to hs37d5 improved the overall variant calling and filtering in the pipeline.

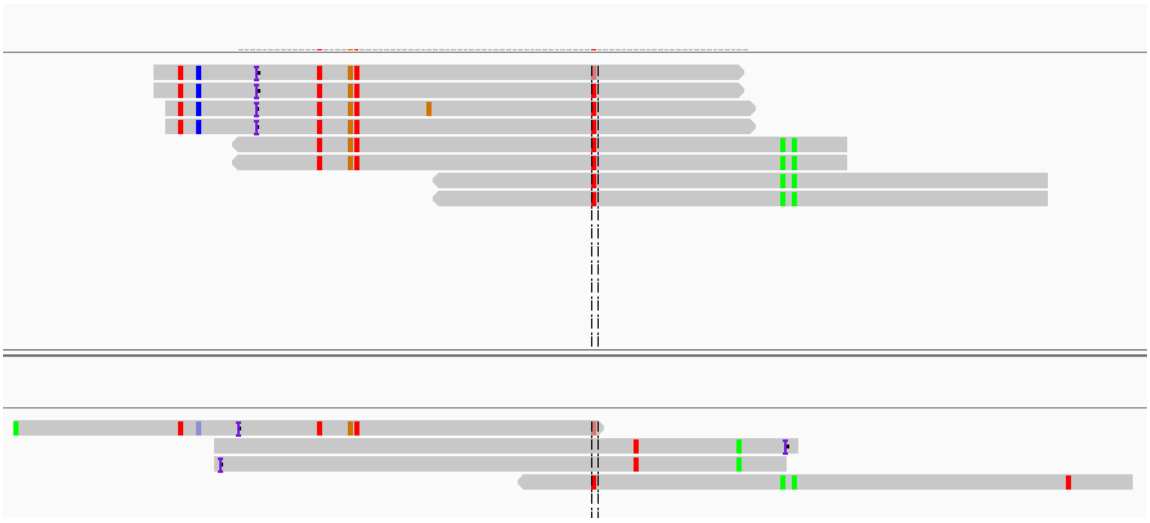


Figure 3.2: Examples of calls on chromosome X unique to Human37 in two unrelated samples.
Shown are noisy WES sequencing reads in IGV on chrX:153,009,900-153,010,200 for two unrelated samples. The grey bars represent individual reads and the coloured bands represent variation from the reference sequence. These reads were not present in hs37d5 as they were all mapped to the decoy chromosome.

3.1.4. Effect of duplicate reads on Platypus variant calling

Since the Platypus variant algorithm is supposed to ignore duplicate reads, it should not be necessary to remove duplicate reads from BAM files before variant calling. To assess if additional provision for duplicate read removal was required for the pipeline, Platypus was run on 10 exome samples before and after duplicate read removal with Picard. The VCFs were run through the Python annotation script and uploaded to IVA. In all 10 cases the total number of variants called by Platypus decreased after duplicate read removal (Table 3.4). There was a small number that passed the confidence filtering that were only called in the VCFs where duplicate reads were retained, of which a very small proportion were deleterious. A selection of these deleterious variants was examined IGV, none of which appeared to be high quality variants and were all in regions with sequencing noise. Since most of the unique variants called in the duplicate read VCFs were low quality, the removal of the duplicate reads with Picard decreased the total amount of probable false

variants in the VCFs. Thus there was a slight improvement in variant calling after duplicate read removal which suggests that Picard was more effective at identifying duplicates. Nevertheless, most of the variants in the duplicate sample VCFs would have been filtered out by the confidence filtering in IVA, a similar result to another assessment of the effect of PCR duplicates on variant calling [100]. The overall decrease of file size by 20% suggests that there was a large amount of superfluous data removed which was not required for downstream analysis. Read counts were assessed using the Python read count script as described in the Methods section 2.5.5. Total mean reads counts of the 38 samples decreased an average of 22%, which was consistent with the overall decrease in file size. The proper computer storage of NGS data is an on going issue [145]. As such, the decrease in data storage requirement associated with duplicate read removal was an important part of the pipeline.

Table 3.4 Comparison of total number of variants called in samples with and without duplicate reads.
Shown are the numbers of variants called in 10 samples before and after duplicate read removal, and the number of high quality variants uniquely called by Platypus in the samples with duplicate reads.

Number of variants when duplicates not removed	Number of variants when duplicates removed	Number of high quality variants uniquely called in samples with duplicate reads	Deleterious
79604	78966	47	3
79744	79179	62	7
77809	77398	45	2
77967	77291	43	5
77073	76674	43	3
76258	75760	38	2
79194	78506	34	10
77444	76901	48	3
80879	80282	39	2
80444	79909	45	7

3.1.5. Exome concordance analysis

To assess the efficacy of the exome sequencing in this project, the genotype concordance of the exome sequencing and cytoSNP-12 array data was determined. Illumina cytoSNP-12 array data had previously been generated for 4 of the samples that were exome sequenced with the SeqCap (NimbleGen) capture kit as part of this project. The overall concordance between the exome data and the genotype calls from the SNP array data was calculated as described in the Methods section 2.5.3. In all four exomes the concordance was over 99% in the exonic regions (Table 3.5). A 'spiked' check was done where the VCF data from one sample was compared with SNP array data from an unrelated sample. The 54% concordance rate in the spiked sample demonstrates that the python script functioned to properly compare the SNP array and exome data.

Table 3.5: Concordance of exome sequencing with cytoSNP-12 array data. Shown is the concordance in 4 samples run with the SeqCap capture kit and on SNP arrays. Spiked sample is the concordance script run on exome data with the wrong SNP array data.

Sample ID	Positions genotyped in exonic target region.	Number discordant	Positions genotyped outside target.	Number discordant
BRC220	2564	4	15506	3303
BRC219	2564	6	16749	3536
1139E1	2558	15	9433	2104
BRC180	2570	7	15828	3102
Spiked	2572	1404	15832	9022

The concordance data was calculated for the first three exomes run with the TruSeq (Illumina) capture kit using Excel (Microsoft) as described in the Methods section 2.5.3. The overall concordance with the SNP array data was 97.4% in BRC_102, 98.0% in BRC_50 and 97.9% in BRC_41. Although the concordance was lower than the SeqCap samples, it was still high and thus not interpreted as an issue with the capture kit. The

data was not run through the python script since it was not necessary to repeat the concordance check. The overall concordance rates in the samples show that the capture kits and WES were effective in providing accurate genotype information. It also showed that there were no sample mix-ups in the exome or SNP array protocols for the 7 samples. It was clear that the concordance within the exonic regions was much higher than outside. This in part validates the approach of only calling variants in the capture kit target regions (see section 3.1.8.2).

3.1.6. Exome contamination and lane merging analysis

The verifyBamID programme was run on all of the exome samples to check that there was no inter-lane sample contamination. The freemix value for the exome samples was within the expected value for uncontaminated samples (Figure 3.3). To create a highly contaminated spiked BAM file, data from two unrelated samples with similar coverage were merged with Picard. Low and moderate contaminated BAM files were created by merging samples with high, medium and low exonic coverage. The --ignore RG option had to be used with the deliberately contaminated BAM files so that the sample names would be ignored from the BAM headers. The freemix values of the contaminated BAM files were as expected, significantly higher than any of the BAM files (Figure 3.3). Since the programme checked the BAM header records, it was also a quality control step to check that none of the BAM files had been incorrectly merged. Since there were no errors stating that the BAM file contained multiple individuals, the raw BAM files were correctly merged. Based on the above analysis, there was no inter-lane sample contamination or BAM files that were incorrectly merged.

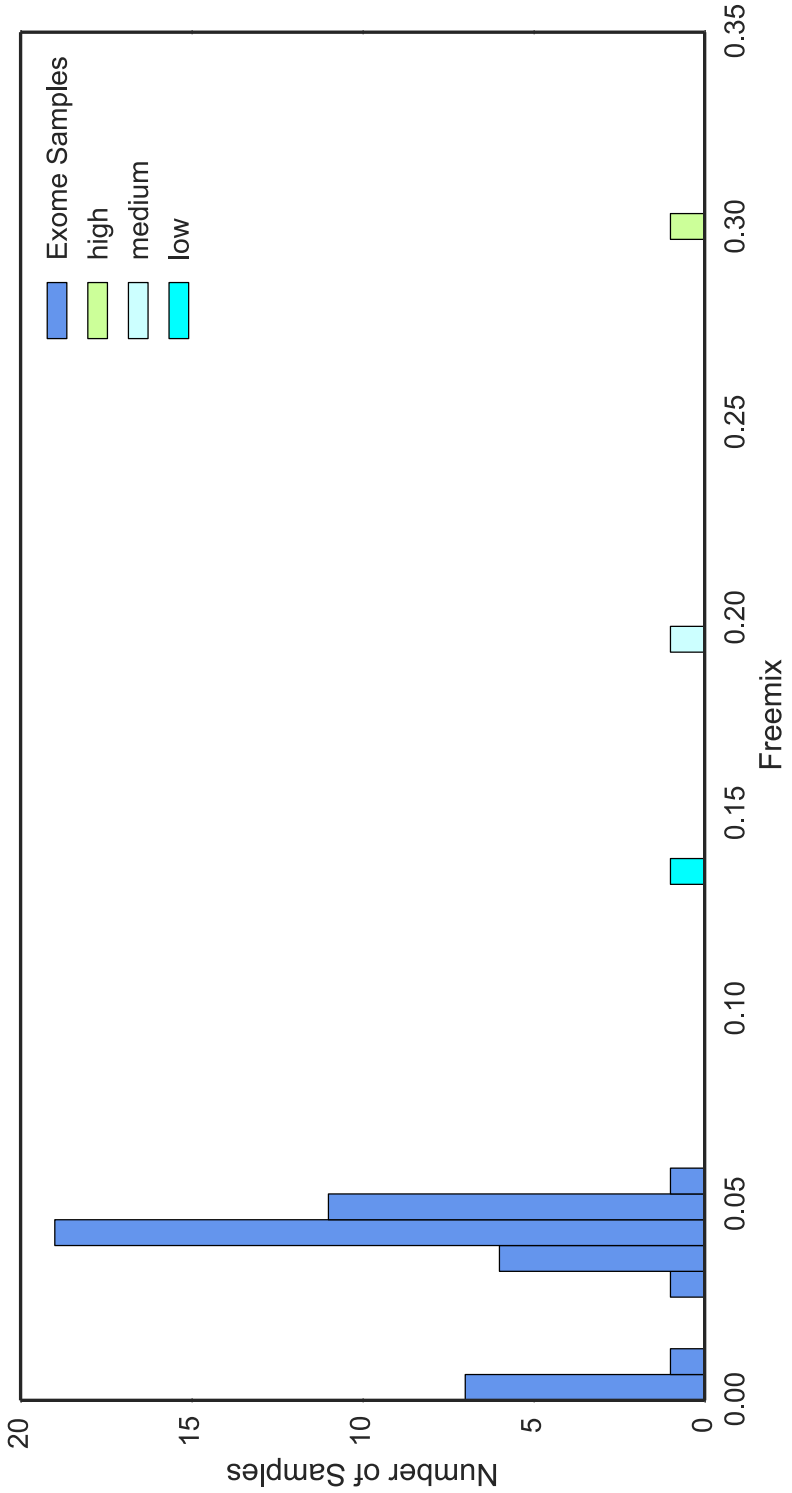


Figure 3.3: Histogram plot of freemix values from verifyBamID programme of exome samples.
Plot shows samples freemix values were much lower than any of the high, medium and low contamination 'spiked' samples. Plot created in ipython notebook with Python packages matplotlib and seaborn

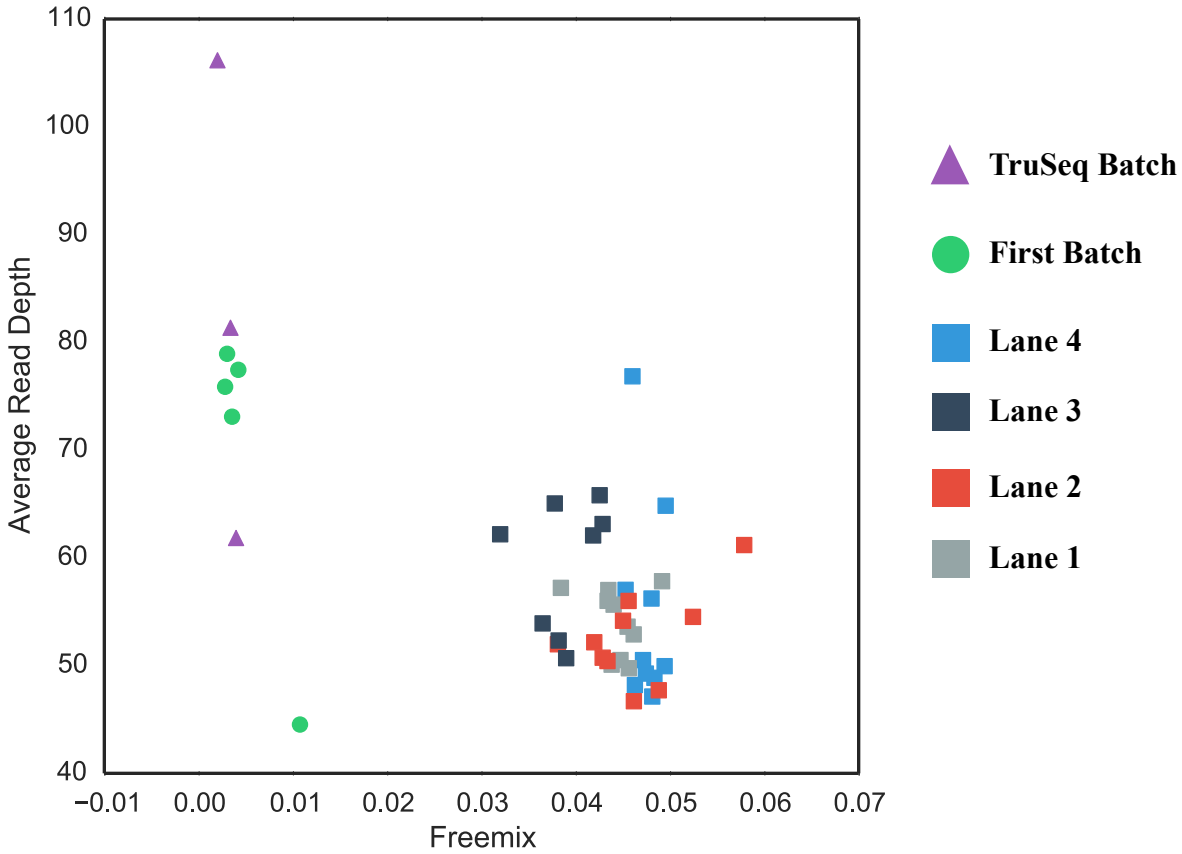


Figure 3.4: Comparison of freemix values between the TruSeq and two batches run with NimbleGen kit.
The plot shows the distribution of Freemix values and read depth from the 6 lanes of sequencing. The initial cohort of exomes run with the SeqCap capture kit (represented by the green circle) have lower overall freemix values compared with the second batch. The three TruSeq samples also had lower freemix values. Plot created in ipython notebook with Python packages matplotlib and seaborn.

A comparison of the exomes run with the TruSeq and the SeqCap capture kits showed that the initial batches performed slightly better based on the freemix metric (Figure 3.4). This could in part be a result of the slightly higher coverage seen in most of the samples from the earlier batches.

3.1.7. Exome read count examination

Sequencing read count is a standard metric for evaluating the overall performance of the sequencing and the capture kit. The mean read depth of the target regions for each sample

was calculated with the tool bedtools coverage and the custom python script as described in the Methods section 2.5.5.

Table 3.6 Validation of python script compared with output from QualiMap v0.8.1.
The output from the python script was compared against QualiMap to ensure that the script was properly calculating the read depth for given regions in a sample (data shown is for sample 1139F1). Listed is the mean read depth for each chromosome, which was concordant between QualiMap and the script.

chr	Python script	QualiMap output	chr	Python script	QualiMap output
1	75.41554569	75.42	14	71.56508328	71.57
2	84.40727762	84.41	15	92.08351831	92.08
3	67.53601795	67.54	16	78.72812942	78.73
4	75.28171423	75.28	17	67.74176543	67.74
5	75.85798508	75.86	18	75.15623339	75.16
6	70.27192554	70.27	19	58.94786202	58.95
7	85.64634556	85.65	20	58.46080572	58.46
8	78.44712149	78.45	21	72.52763249	72.53
9	83.62202187	83.62	22	71.48595591	71.49
10	85.11645867	85.12	X	111.3560444	111.36
11	67.55197069	67.55	Y	2.933615047	2.93
12	68.91415692	68.91	MT	362.333957	362.33
13	74.91888975	74.92			
Mean total coverage calculated					
	75.86992416	75.87			

The mean read depth per chromosome calculated with the Python script was concordant with the output from Qualimap v0.8.1 (Table 3.6 and Figure 3.5), which validated the Python script. The average mean read depth of the target regions for all the samples was 56.74X (not including two failed samples), which exceeded the objective for this project of between 25 to 30X (Table 3.7 and Table 3.8). The total percentage of the exome with greater than 10X coverage was 90%. Two of the parental samples, BRC_153F and BRC_153M, did not perform well on either lane of the sequencing (Table 3.9).

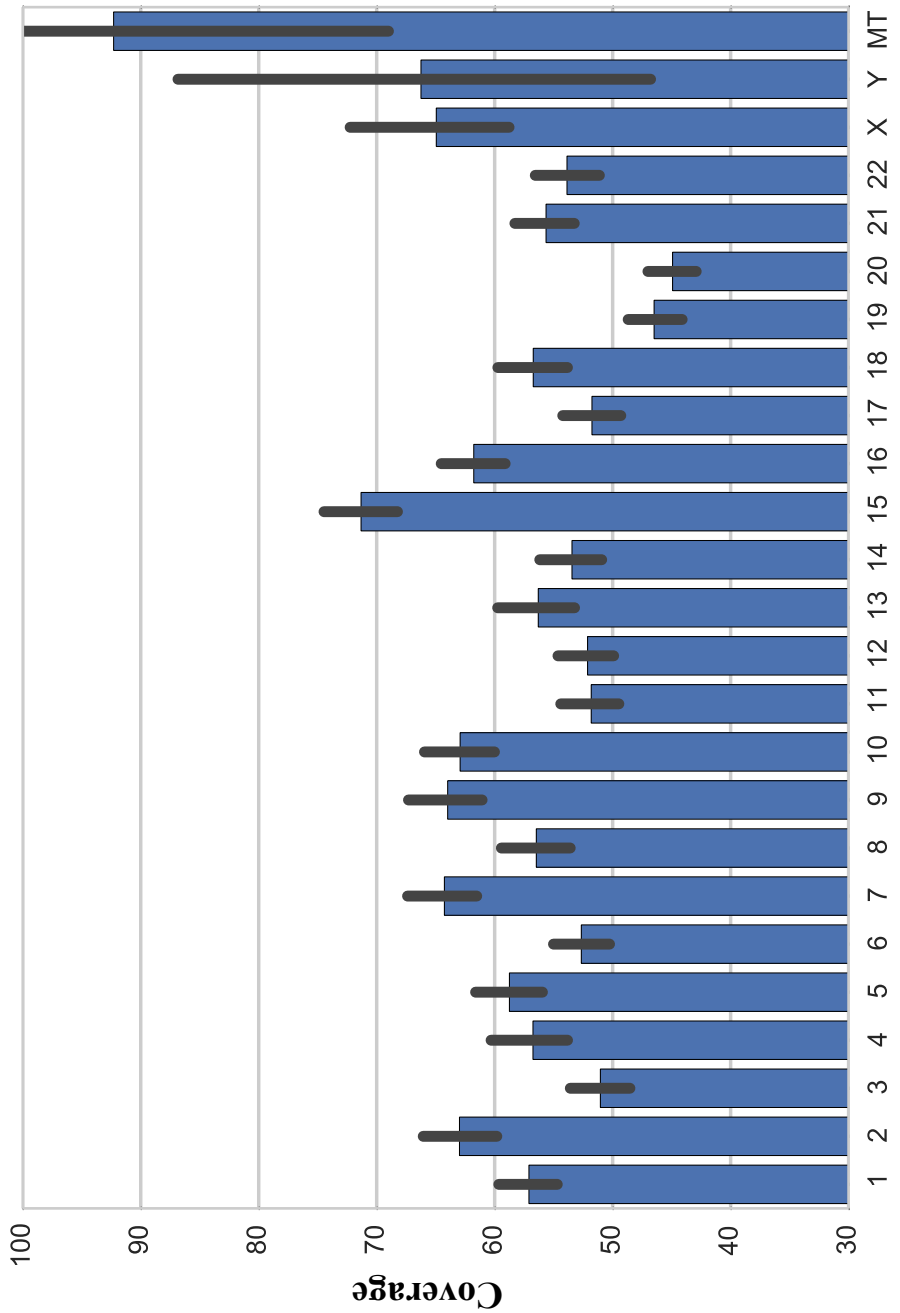


Figure 3.5: Mean target read depth for each chromosomes for samples run with SeqCap (NimbleGen) capture kit.
Data is shown for the mean chromosome read depth in 43 samples (not including failed parental samples) run with NimbleGen capture kit and calculated with the custom Python read depth script. Error bars are the standard deviation. Plot created in ipython notebook with Python packages matplotlib and seaborn.

Table 3.7: Mean and median depth of coverage for 43 samples run with NimbleGen capture kit and percentage covered at 10X, 20X and 30X.

Shown is the mean and median read depth calculated for SeqCap V3 target regions with the custom Python script. The total percentage of the target regions covered with at least 1 read is shown along with the percent at 10X, 20X, and 30X coverage. (failed parental samples are not included, see Table 3.9).

	Coverage		% target	Percent covered at		
Sample ID	Mean	Median	covered	10X	20X	30X
1139E1	44.53	37	95.63	88.64	77.08	60.98
1139F1	75.87	63	96.06	92.38	88.42	82.12
BRC_115	57.83	46	96.49	91.09	83.11	71.36
BRC_115F	62.08	50	96.63	91.83	85.04	74.74
BRC_115M	52.33	43	96.40	90.55	81.42	68.07
BRC_146	62.18	51	96.67	92.03	85.47	75.42
BRC_146F	53.91	44	96.92	91.61	82.91	69.87
BRC_146M	65.04	54	97.06	92.63	86.73	77.54
BRC_153	50.68	41	96.73	90.90	81.30	67.11
BRC_16	49.26	40	95.19	88.66	78.99	64.87
BRC_16F	48.19	39	95.03	88.15	77.87	63.31
BRC_16M	57.03	46	95.77	90.54	83.19	71.86
BRC_224	65.80	54	96.52	91.98	86.21	77.30
BRC_225	63.13	51	96.69	91.81	85.20	75.26
BRC_225F	52.16	42	96.74	90.90	81.26	67.42
BRC_225M	61.18	50	96.71	91.84	85.06	74.70
BRC_226	46.69	38	96.89	90.38	78.54	62.14
BRC_228	47.71	39	96.69	90.63	79.33	63.39
BRC_230	51.98	42	96.66	90.95	81.65	68.01
BRC_230F	50.42	41	96.82	90.89	80.62	65.89
BRC_230M	54.16	44	96.86	91.39	82.78	69.94
BRC_231	54.52	44	96.90	91.04	81.80	68.68
BRC_232	56.00	45	96.98	91.81	83.52	71.02
BRC_232M	50.73	40	96.74	90.14	79.28	64.77
BRC_59	56.22	46	95.83	90.51	82.90	71.15
BRC_59F	49.94	40	95.69	89.72	80.03	65.62
BRC_59M	48.86	40	95.43	89.08	79.19	64.70
BRC_81	50.51	41	95.66	89.61	79.93	65.71
BRC_81F	47.13	38	95.42	88.79	77.81	62.20
BRC_81M	76.84	62	95.93	91.51	86.86	80.19
BRC_83	64.83	52	96.98	91.27	85.15	75.88
BRC_83F	53.61	43	96.47	90.77	81.81	68.78
BRC_83M	50.09	41	96.53	90.18	79.99	65.57
BRC_84	57.21	46	96.65	91.44	83.52	71.70
BRC_90	57.01	46	96.79	91.50	83.68	71.92
BRC_90F	56.01	45	96.80	91.41	83.02	70.60
BRC_90M	49.75	40	96.55	90.35	80.20	65.65
BRC_91	52.89	42	96.70	90.73	81.07	67.40
BRC_91F	50.51	41	96.54	90.48	80.33	65.87
BRC_91M	55.65	45	96.32	90.79	82.63	70.51
BRC180	77.45	64	96.32	92.85	89.13	83.14
BRC219	78.93	65	95.85	92.02	88.10	82.21
BRC220	73.09	61	95.82	91.83	87.39	80.53

Table 3.8: Mean and median coverage of three samples run with TruSeq capture kit and percentage covered at 10X, 20X and 30X.
Shown is the mean and median read depth calculated for the TruSeq target regions with the custom Python script. The total percentage of the target regions covered with at least 1 read is shown along with the percent at 10X, 20X, and 30X coverage.

	Coverage		% target	Percent covered at		
Sample ID	**Mean**	**Median**	**covered**	**10X**	**20X**	**30X**
BRC_102	106.17	89	97.80	93.55	90.18	86.32
BRC_50	61.84	55	96.95	90.67	84.77	76.78
BRC_41	81.35	71	97.45	92.31	87.82	82.44

Table 3.9: Read counts of failed parental samples BRC_153F and BRC_153M
Shown is the mean and median read depth calculated for SeqCap V3 target regions with the custom Python script for the two sets of sequencing of the parental samples BRC_153F and BRC_153M. The total percentage of the target regions covered with at least 1 read is shown along with the percent at 10X, 20X, and 30X coverage.

				% target	Percent covered at		
	Sample ID	**Mean**	**Median**	**covered**	**10X**	**20X**	**30X**
First Run	BRC_153F	17.22	14	95.01	63.11	28.48	11.14
	BRC_153M	15.46	12	94.35	57.37	23.53	8.72
Repeated Sequencing	BRC_153F	69.23	6	65.28	47.17	43.78	41.38
	BRC_153M	61.00	5	63.68	47.05	43.51	40.97

The libraries of the two parental samples were re-sequenced on a single lane, in a barcoded batch of two samples. The initial quality control step performed by the Core on the data did not highlight any issues as the exonic mean coverage was 91.07X and 81.15X and the median coverage was 70X and 65X. The data was downloaded from the Core servers and run through the pipeline to remove duplicate reads. A visual inspection of the BAM files in IGV initially highlighted that there was still a significant issue with the data (Figure 3.6). It appeared that despite the apparent high mean and average coverage, certain genes were not covered. The BAM files were run through the read count pipeline and assessed with the custom python script. After duplicate read removal, the average read count was 69X and 61X, but alarmingly the median read count was 6X

and 5X. The BAM file size decreased 42% and 43% after duplicate read removal, which suggested there were significant issues with PCR artefacts in the sequencing. The issue with the data became clear when the specific coverage of genes was assessed.

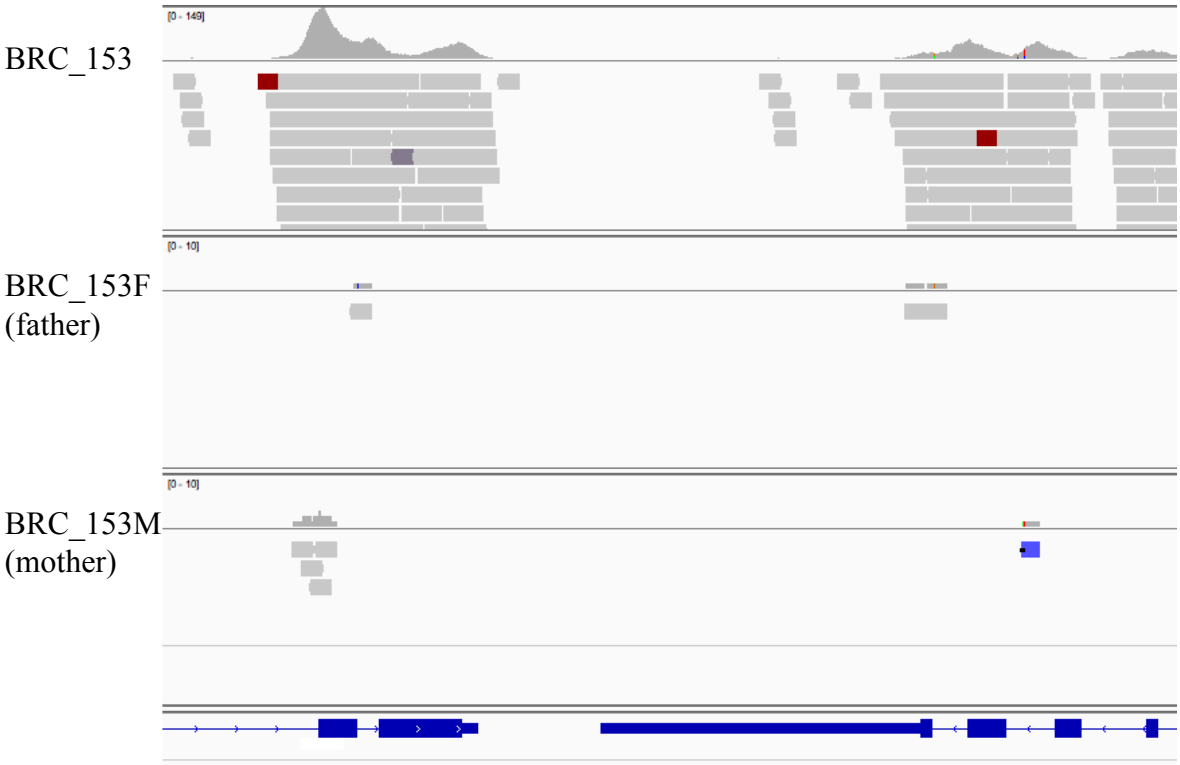


Figure 3.6: IGV visualization of BRC_153 coverage compared with parental samples.
Figure shows the visualization of a region in IGV where the sequencing of BRC_153F and BRC_153M failed despite the apparently high mean coverage. The grey bars represent exome sequencing reads aligned to exonic regions (blue bars at the bottom). In this exonic region, BRC_153 had sufficient read depth for variant calling, yet there was barely any coverage in the parental samples highlighting that the sequencing failed.

3.1.7.1. Read count assessment of genes of interest

The custom python script was designed to check the read coverage and highlight any genes of interest that had low mean read depth or no coverage at all. Specifically, the mean read count of all 436 genes associated with brain abnormalities (Appendix Table 1)

from the Ingenuity Knowledge Base was checked. All genes of interest had coverage (mean read depth > 0), except for BRC_153F and BRC_153M where the sequencing failed for several genes. The mean read depth of *FGF22* and *IL17D* was less than 4 in five samples. Connection of both genes to brain abnormalities was examined and not found to be directly linked to any disorders. The association of *FGF22* was based on a study that showed that the protein FGF22 was required for normal midbrain development in zebrafish [146]. The IVA knowledge base associated *IL17D* with brain abnormalities since it was one of 15 regulators of the proinflammatory cytokine pathway which is involved in brain edemas [147]. Since neither gene had previously been directly implicated in a brain abnormality, it was not considered a significant issue that the coverage was quite low in those genes.

Table 3.10: Brain abnormality genes with less than 10X coverage in samples.
Total number of samples in the cohort of 46 with low coverage listed.

Gene	Samples with less than 10x	Mean coverage in low samples	Brain Abnormality Phenotype
ARX	20	7.41	EIEE, Lissencephaly, MR
CDT1	3	8.43	Meir-Gorlin Syndrome
CYP26C1	36	8.21	Microcephaly mouse model [148]
FOXH1	1	9.1	Holoprosencephaly [149]
LMNB2	1	9.81	Microcephaly mouse model [150]
No direct association			
DVL1	5	8.31	
FGF22	41	6.10	
GAS1	10	7.71	
IL11	7	8.40	
IL17D	38	7.41	
MESP1	20	8.28	
MIB2	17	7.02	
PURA	7	8.79	
SOD3	16	9.07	
BAT3	1	9.23	
TNF	1	7.73	

Since less than 4X coverage is quite low for accurate variant identification, all of the genes of interest with less than 10X coverage were also assessed (Table 3.10). There were 16 genes of interest, of which 5 were directly associated with a brain abnormality phenotype. This was a similar result to other studies of the variable coverage of WES in genes of interest [5].

The Python read count script functioned well to highlight the issues with the two parental samples that failed. It showed that in each sample, there were genes of interest that had no coverage at all. There were also a significant number of genes of interest with less than 4X coverage, 124 in BRC_153F and 128 in BRC_153M, which showed that the original sequencing of the samples had performed better in terms of genes of interest. There were several reasons why the sequencing could have performed badly, including sample and library quality. Whatever the cause, the data from the re-sequencing was of such low quality that the VCFs from the data were only analyzed briefly to see if there was anything of note. The quality data for these samples was reported back to the Core sequencing group and the parental samples will be sequenced in the future with a new DNA aliquot. The low quality data did function to highlight the important of a proper quality control step run on the BAM files. The Python script was particularly useful since it highlighted genes of interest with low coverage.

The QC assessment of the data showed that the WES was in most cases of high quality. The two cases where the data was problematic were quickly identified. The read depth assessment of the samples run with the SeqCap and TruSeq capture kits show that WES

functioned well to capture the majority of brain abnormality genes of interest for this project. The 16 genes of interest with less than 10X coverage were further evidence that WES is not yet a perfect technology. As other studies have found, WGS could be a better tool due to its uniform genomic coverage and increased sensitivity [16].

3.1.8. Platypus variant caller

During the course of the project Platypus was updated to version 0.5.2. The variant calling for the first three WES samples was done with Platypus 0.1.8 by the Bioinformatics Core service at the WTCHG. The subsequent WES samples were run on Platypus 0.5.2 as part of the research project.

3.1.8.1. Variant calling and annotation issue with Platypus minFlank setting

An issue was encountered with the new version of Platypus when it was run with the default settings. Neighbouring variants on the same haplotype were being called as a single variant and as a result common variants were not being correctly annotated. MinFlank (mF) settings for Platypus control the way in which variants on the same sequencing read are reported. If two neighbouring variants are *in cis* and within the default setting of 10bp, Platypus will call it as a single in frame substitution. An example of the problem with default mF settings was in the polymicrogyria (PMG) sample 1139F1. In the first version of the VCF file, an inframe substitution was called in the known microcephaly gene *ASPM* (Table 3.11).

Table 3.11 Example of two common variants in *ASPM* called as a single variant by Platypus.
The table shows the output from Platypus where two common SNPs (underlined) are called as a single variant.

#CHROM	POS	REF	ALT
1	197070697	TTTGTATTACG	CTTGTATTACA

Platypus called the variant as an 11bp substitution, which IVA was not able to annotate relative to known variation and simply called it a large in frame insertion. The insertion was actually two variants, c.7674C>T and c.7684A>G (Figure 3.7). Both are common dbSNP variants, which are present in ClinVar with the classification of no known pathogenicity. To check that the annotation issue was not an error with Ingenuity's pipeline, the variant calls for the *ASPM* example were run through the command line annotation tool VCFtools vcf-annotate [151] which was also not able to annotate the large insertion variant relative to dbSNP138.

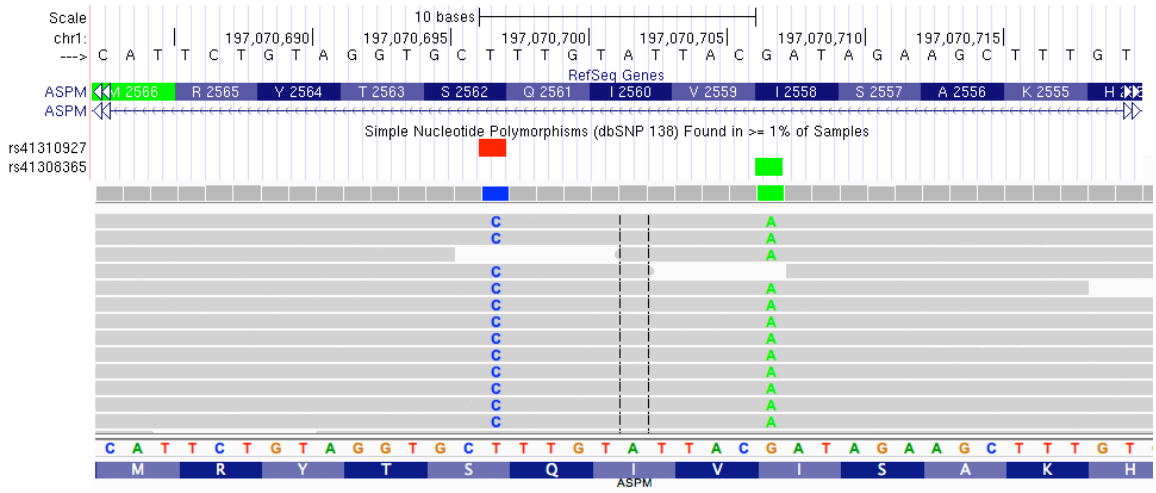


Figure 3.7: Example of the two common variants in *ASPM* shown in IGV and in UCSC Genome Browser.
The exome sequencing reads are shown in IGV, with the variant T > C and G > A are aligned with the common dbSNP track in UCSC Genome Browser. Both SNPs are common variation at >1% frequency.

The *ASPM* variant was an example of the issue with the mF default setting; common variants in known brain abnormality genes were not filtered out by IVA and appeared as novel in frame substitutions. It was not until the region in question was examined in UCSC Genome Browser that the mis-annotation was discovered. If this issue only applied to a small number of variants it would not have been difficult to examine each in UCSC and exclude them as candidates. Unfortunately there was a large amount of common variants on the same allele within 10bp.

To examine the common variants that were mis-annotated and compare the overall difference between the mF default settings and mF set to zero, two versions of a VCF were made with Platypus from the same BAM input file with matching settings except for the mF flag.

Table 3.12: Comparison of variant calls with different minFlank settings.
The total number of variants called in a sample with two different minFlank settings are listed along with the number of variants that were not annotated by IVA.

minFlank setting	Number of dbSNP variants	Number of dbSNP variants that were not annotated
10 (default)	67257	6596
0	73185	341

Two case-control analyses were created in IVA with each VCF as control and the other as the case. A simple filtering was used to examine the variants called by one set of settings but not the other (Table 3.12). It was clear from the 6596 dbSNP variants only reported in the mF=0 VCF that the majority of common variants present on the same allele were annotated as separate events in dbSNP. Only a small number of multi nucleotide insertions, 341, in the mF=10 VCF were present in dbSNP. Intermediate

options were considered as well with mF set to 1, 2 and 3 but a similar result was produced to that described above.

In Platypus version 0.2.0 and earlier, the default mF setting was 3 and there was no issue with mis-annotated variants. In a comparison of a sample called with Platypus 0.5.2 and 0.2.0 with mF set to 3 in both versions, close to 10% of the dbSNP variants present in the 0.2.0 VCF were not annotated correctly in the 0.5.2 VCF. Despite having the same mF option, it was clear that the mF setting was interpreted differently in the newer versions of Platypus. After extensive personal communication with Dr Andy Rimmer, the Platypus programmer, it was suggested that the mF set at 0 should produce a similar result to 0.2.0 and fix the annotation issue. It was not considered a viable option to use the earlier version of Platypus since significant improvements had been made to the calling algorithm in the newer versions. A subsequent issue was encountered when looking at the distribution of variation in VCFs that were created with the newest version of Platypus with the mF set to 0. The VCF had no large substitution calls since each event was called separately (Figure 3.8). The main concern with using the mF set to 0 option was that the large substitution calls that were not common variants were not being identified as they were in Platypus version 0.2.0.

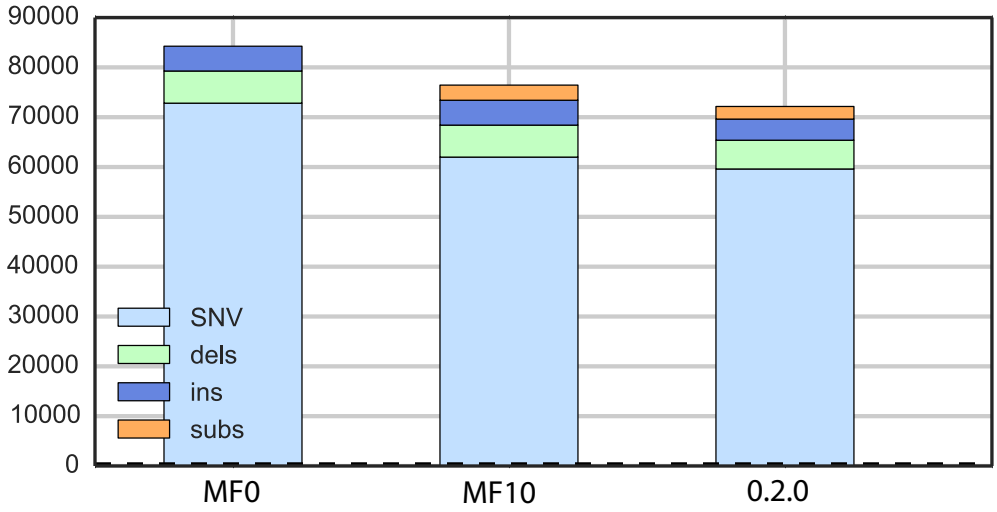


Figure 3.8: Plot showing the distribution of variation between mF=0, mF=10, and Platypus 0.2.0 VCFs.
The stacked bar chart shows the distribution of variation type with different versions of Platypus and minFlank settings. The minF 0 VCF had no large substitution calls (subs) compared with the other VCFs, which was a concern. The plot was created in iPython notebook with matplotlib and seaborn.

The process of filtering common variation out of sequencing data is a critical step for variant discovery. Since it was not possible to annotate the majority of common variation with the default mF option, it was set to 0 for the variant calling in this project. Although this was not the perfect solution for variant calling, it represented the best solution for downstream variant annotation in Ingenuity Variant Analysis. For the purpose of filtering out common variation from sequencing data it was important that the variant calling closely matched that present in the existing databases. The mF issue was not unique to this project. A solution being developed is a script that will split the variants called by Platypus to allow for annotation. Unfortunately, the programming that the script entails was beyond the scope of this project. Significant feedback in regards to the mF issue was given to Andy Rimmer and hopefully a solution will be found either with a new version of Platypus or better annotation pipelines.

3.1.8.2. Result of only calling variants in the exome kit target regions with Platypus

The overall quality of the data outside of the target regions was assessed to see if high quality variants were missed by only calling variants in the NimbleGen target regions. Concordance of variants with array data outside of the exonic regions was less than 80% compared with over 99% inside the regions (Table 3.5). When comparing four samples called with and without the region file, the total number of unfiltered variants decreased on average from ~300,000 to ~90 000. The average Phred scaled probability scores (QUAL) for the variants from an exome VCF called on the whole genome was ~250 outside of the exome regions versus ~1000 in the regions. The decrease in total calls facilitated filtering and identification of candidates from the data. There were likely some real variants captured outside of the target regions but based on the concordance outside of the target regions, the number of false variants would severely hamper the analysis. For these reasons, only the exonic target regions of the capture kits were used for the variant calling regions in this project.

3.1.8.3. Validating the inclusion of Mitochondrial Genome as part of target regions for variant calling with Platypus

Since there were no mtDNA targets in the WES kits all of the mtDNA coverage was from off target sequencing reads. To assess if the mtDNA calls were high quality with exome sequencing, the mtDNA calls from three samples with both WGS and WES were examined. Unfortunately there were no WES samples run with the Nimblegen SeqCap available with corresponding WGS data. A previous study on a single sample had found that the mtDNA sequences derived from three different WES capture kits, SureSelect (Agilent), TruSeq and SeqCap were all identical and concordant with WGS [115]. Since

there was probable consistency between the WES kits, the analysis was done despite the fact that the WES was done with a different capture kit (SureSelect) from the current project.

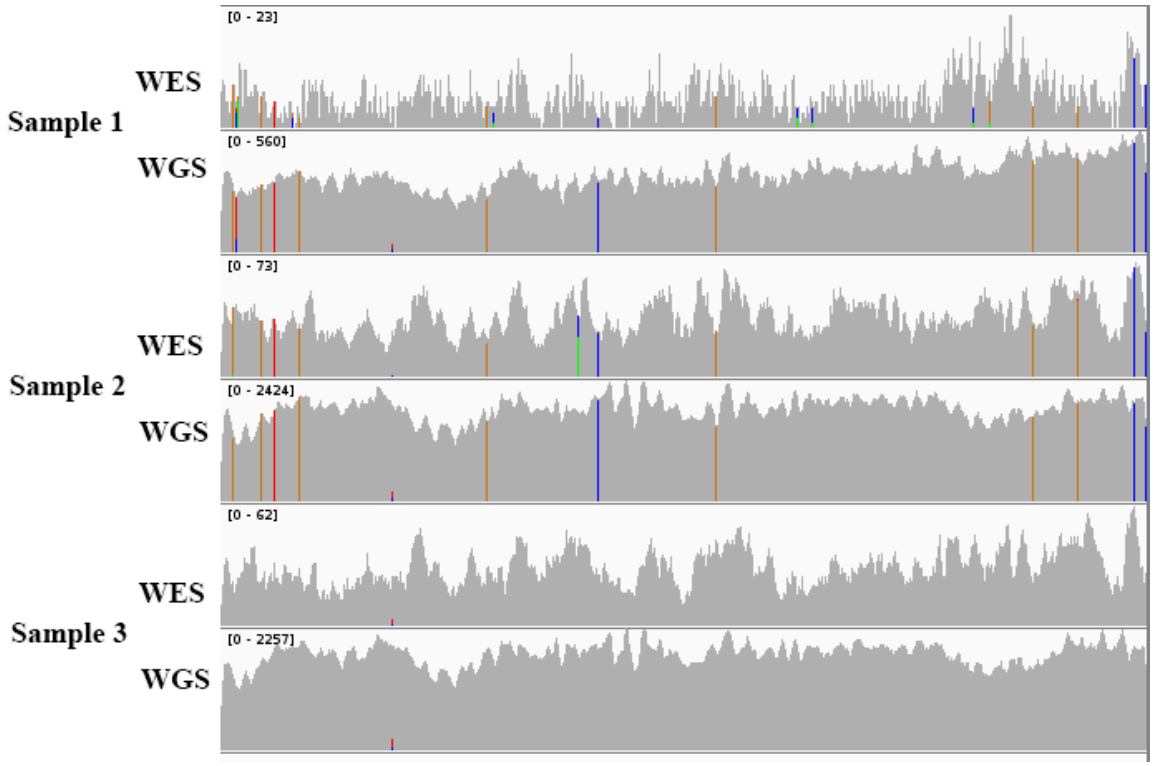


Figure 3.9: WGS and WES comparison of mitochondrial regions visualized in IGV.
Three samples with corresponding WES and WGS data are shown in IGV. The grey represents the total sequencing coverage across the mitochondrial genome. Coloured lines represent variant positions, which visually appear to be concordant between WES and WGS in sample 1, 2 and 3.

A visual inspection of the BAM files in IGV showed concordance between WES and WGS (Figure 3.9). To examine the specific mtDNA calls, Platypus was run targeting only the mtDNA region on the three WES and three WGS BAM files.

Table 3.13: Number of concordant variants called in mitochondrial regions in WGS compared with WES.
Listed are the total variants called in the mtDNA regions in the WGS, WES and the total number concordant.

	Genome	Exome	Concordant
Sample 1	17	9	9
Sample 2	14	15	12
Sample 3	3	2	1

In all of the samples, there were 26 variant calls in the exome data of which 22 were concordant with the genome data (Table 3.13). The additional variants in the genome data were missed due to regions of low coverage in the exome. Non-concordant variants in the exome data were heterozygous with low read support (<10% MAF) and would not be a source of false variants, as they would be directly filtered out by IVA. Although some variants were missed in the WES data, all of the high quality variants (QUAL > 20) called were present in the genome data suggesting that the mitochondrial regions were effectively covered by WES. Based on this analysis, including the mtDNA genome in the exome target regions for this project would not increase the number of false variants called in the exome and may have helped detect variants of clinical relevance.

3.1.8.4. Examination of Platypus multi-sample VCFs in trios and families

Running Platypus as a single instance on all family members has been reported to improve variant discovery and *de novo* variant filtering [107].To ascertain the effect of the improved variant calling in this pipeline, the variant calls for two parental samples were compared between the multi-sample family VCF and the VCFs where Platypus was only run on the individual sample. The samples were run through the standard pipeline and uploaded to IVA. There were 11.8% and 11.7% more variants called in each individual in the multi-sample VCF compared with their individual VCF. When these

variants were examined roughly 25% of them in each case were dbSNP variants. This suggested that a good portion of those variant calls from the multi-sample VCF were likely real and not called in the individual VCF due to insufficient coverage. Of those variants uniquely called in the multi-sample VCFs, roughly 12% of them were predicted as missense changes. Since there was an increase in variant calls that appeared to be real, Platypus was run on whole families when available for variant calling in this project. The increase in variants calls due to calling the samples together likely improved the downstream filtering for *de novo* variants in families.

3.1.9. Validation of Ingenuity Variant Analysis for filtering and analysing sequence data

Since IVA was not part of the standard WTCHG bioinformatics pipeline, the efficacy of the software had to be assessed. The initial validation analysis focused on seven Early Infantile Epileptic Encephalopathy (EIEE) families from the WGS500 project. Six of the families were recently published with strong candidate genes [94]. The samples were initially uploaded to IVA using the web browser based upload tool. This method was problematic as the upload tool crashed several times when attempting to upload WGS data. As a result, an FTP upload account was setup with IVA and the files were uploaded directly from the servers with the UNIX curl programme.

The first validation analysis was to compare the total number of variants reported in IVA with the raw number of variants in the VCF files. The total number of variants reported by IVA in all 22 EIEE samples matched the number of variants expected from the raw VCF. In a more in depth examination of one of the samples, the exact number of variants

reported per chromosome was checked and found to be concordant. As a second check, the number of calls reported in a single large gene *TTN* was also compared in the same sample and was also found to be concordant. Although FTP is a well-established method for transferring files [152], the variant count check also served to ensure that there was no data loss when uploading files by this method.

The data from the seven EIEE families were initially analyzed together as one whole group without any prior knowledge of the candidate variants. The only biological context provided for the filtering was "Ohtahara", another term for EIEE. A causative variant was identified in the known EIEE gene *KCNT1* with the default IVA settings. After the initial analysis, each family was analysed independently. In 3 of the 6 published families, the causative variants in *KCNT1, SCN2A,* and *KCNQ2* were identified in the first pass analysis. The *PIGQ* variant, present in one family, was found once the biological context was expanded from Ohtahara to also include seizures. This was a valuable lesson about the IVA knowledge base and emphasized the importance of expanding the biological context with key phenotypic terms if no strong candidates were apparent. Candidate variants in the final two published EIEE families were not initially identified, as the biological connection in the IVA knowledge base with EIEE was distant. In a retrospective analysis of the six EIEE families, three months after the paper which described the association of the genes with epilepsy was published [94], all of the variants were listed as directly linked to Ohtahara syndrome. This demonstrated that the IVA knowledge base is updated regularly with new genes, at least quarterly.

To further test IVA, several WGS500 samples without strong candidates were extensively analyzed. These cases included the unpublished EIEE family quartet, two unrelated renal disease families, and two Lipodystrophy families. This data has already been analysed by WTCHG bioinformaticians looking for variants of likely pathogenicity and consistent with this, in most of cases no candidates were found. A complex indel variant was identified in *UMOD*, a known gene for kidney disorders in the Familial Interstitial Nephritis case. The previous analyses had focused on a large linkage region on chromosome 16 based on SNP array data from the large family pedigree of 26 individuals. The variant, p.V93_G97del/ins AASC, was subsequently validated by Sanger sequencing and had previously been implicated in kidney disorder with supporting functional studies and protein modelling [153,154]. The variant was 4Mb outside of the linkage region and the linkage analysis failed because there was a second cause of renal failure in the family. This case highlighted the strengths of IVA. It was relatively straightforward to first analyze the linkage region and subsequently expand the analysis to the whole genome where a candidate was identified.

A validation analysis was done on the variant annotation pathway in IVA. A VCF containing only dbSNP variants was uploaded to IVA through the standard pipeline. In IVA, 99.4% of the variants were annotated as present in dbSNP. Only a small number of variants were not annotated with a dbSNP number and these were all multi-allelic variants. The final validation examination of the pipeline was on the specific annotation assigned to variants. A VCF was uploaded to IVA and the raw file was also annotated with the command line Variant Effect Predictor (VEP) programme [155]. The VEP

programme determines the consequence of variants on protein sequence and was part of the standard WTCHG bioinformatics annotation pipeline. To examine the specific annotation, a selection of variants, at least 2 from each chromosome were examined in IVA and compared with the VEP calls in the annotated VCF. In all cases the annotation in IVA matched one of the annotations listed by VEP. Occasionally VEP would prioritize a different transcript or gene to list first but the second annotation matched IVA. The SIFT and Polyphen scores were also concordant in the variants examined.

The validation analyses run on IVA showed that the tool accurately imported and interpreted data from VCF. Although the tool differed in how it annotated and interpreted certain variants, these differences would not significantly affect analyses. There were several other bioinformatics VCF tools available [103,151,155,156], but none of them provide the same package of filtering and annotation tools present in IVA. The most powerful of which was the Biological Context Filter, which provided access to the proprietary Ingenuity Knowledge Base. Finally, IVA offered an online graphical user interface which was more user-friendly than many of the command line based tools. For all of these reasons, Ingenuity Variant Analysis was used as the main bioinformatics VCF filtering tool for the WES data in this project.

3.1.9.1. Validation of python script NR_to_DP_AD_VF.py for modifying VCFs for IVA

The original pipeline for annotating and uploading VCF samples to IVA involved creating a multi-sample VCF which was subsequently split by columns into individual samples using the UNIX cut command and modified with the original Perl script. It

successfully split the VCF but incorrectly kept the same VCF header for each file. A potential fix for the error was modifying each VCF header but this would have been a further workaround to keep using the original Perl modification script that had other inbuilt errors.

Since a new python script was written to modify the VCFs for Ingenuity, it required validation. Three different versions of a VCF were created to ascertain the functionality of the VCF modification scripts. Firstly, the unmodified output from Platypus was uploaded directly to Ingenuity. As expected the read depth and variant frequencies fields were not present as Ingenuity does not recognize non standard VCF annotation flags as defined by the 1000 Genome project. A modified version of the VCF was created with the Perl script vcf_format_AD_DP.pl written by Gareth Wilson and finally a third VCF version was created with the custom Python script NR_to_DP_AD_VF.py. Both of the modified VCFs were uploaded to Ingenuity and analyzed. As expected, the total number of variants reported in each sample by Ingenuity was not altered by any of the scripts.

The two scripts were concordant for read depth (DP) and depth per allele (AD) at the majority of positions. Of the 84248 variants, the two scripts differed at 1268 positions, all of which were multi-allelic variants.

Table 3.14: Comparison of Perl and Python script modified genotype field outputs.

Unmodified Platypus genotype fields output	
GT:GL:GOF:GQ:NR:NV	2/1:-1,-1,-1:17:3:66:8,56
Perl script output	
GT:GL:GOF:GQ:NR:NV:AD:DP:VF	2/1:-1,-1,-1:17:3:66:8,56:58,8:66:0.1212
Python script output	
GT:GL:GOF:GQ:NR:NV:AD:DP:VF	2/1:-1,-1,-1:17:3:66:8,56:2,8,56:66:0.9655

The Perl script had an error where it interpreted multi-allelic values incorrectly. The Perl script would select NV values as a string and selected in certain cases the smaller of the two values. An example of a multi allelic variant with NV= (8 , 56) and NR = (66) is shown in Table 3.14. The Perl script selected 8 as the variant number and ignored the other variant, since it interpreted the maximum of a string alphabetically, where 8 comes ahead of 56. As a result the AD was reported incorrectly as 8 variant and 58 reference positions with Ingenuity calculating a VF of 12%. The Perl script calculated the number of reference alleles incorrectly since it did not subtract the number alternate alleles for both variants from the total read depth. In this example, the actual depth per allele at this position was 2 reference, 8 one variant and 56 the other variant as correctly calculated by the Python script.

Based on these analyses, the script NR_to_DP_AD_VF.py modified multi sample VCFs correctly for the read depth fields to be imported into IVA. The script was a standard part of the pipeline for variant identification and prioritization in this project.

3.2. Identification of candidate variants associated with brain abnormality phenotypes

After the validation of the pipeline for analysing WES data, I applied it to a large cohort of patients. The following section focuses on the variants of interest identified in the cohort. Of particular focus are the clinically relevant variants and the novel disease gene discovered.

3.2.1. Overview of cohort

In this project, 23 patients with brain abnormality phenotypes were exome sequenced along with 25 unaffected family members. The cohort included a range of brain abnormality phenotypes with most patients presenting with microcephaly. The VCF data produced from the exome sequencing was analyzed with Ingenuity Variant Analysis as described in the Methods section 2.7.5. Candidate variants were identified in 16 of the 18 families that were successfully sequenced. The results of all families are summarized in the table (Table 3.15). Clinically relevant findings were identified in 5 of the 18 families giving a diagnostic rate of 27%. Based on the functional work conducted and recent publications, 2 further results are likely to be pathogenic for an overall diagnosis rate of 38%. Clinically relevant variants identified are summarized followed by an examination of the novel variants of interest found in *PGAP3*, *PI4KA* and *SOX11*. The *PGAP3* result is explained in depth, as it is the exemplar result of this project.

Table 3.15: Candidate brain abnormality variants identified by WES

Variants listed in the table were identified with IVA using the filtering strategy describes in Methods 2.7.3 and 2.7.5. The inheritance pattern listed was established based on the candidate variants identified or the reported consanguinity. All variants listed were validated by Sanger sequencing (table continues on next page).

Patient ID	Patient / Family Phenotype Summary	Gender	Established inheritance pattern	Gene	Coordinates	EVS freq	Protein Variant	Transcript Variant	Transcript ID	Translation Impact
Clinically Relevant findings										
BRC_115	Relative microcephaly, seizures and significant developmental delay	F	*de novo*	*PCDH19*	chrX:99662577	0.00%	p.N340S	c.1019A>G	NM_001105243.1	Missense
BRC_16	Microcephaly (-5SD), cerebellar hemisphere hypoplasia and speech issues.	F	*de novo*	*CASK*	chrX:41420814	0.00%	p.M549fs*2	c.1647_1648delGC	NM_001126054.2	frameshift
BRC_59	Microcephaly (-5SD to -6SD), seizures, and hypoplastic cerebellum and pons.	F	*de novo*	*CASK*	chrX:41420836	0.00%	p.T542fs*46	c.1623_1626delAACA	NM_001126054.2	frameshift
BRC_81	-5 to -6SD microcephaly, cortical dysplasia	M	compound HET	*CDK5RAP2*	chr9:123171404	0.00%	Splice donor loss	c.4604+1G>C	NM_018249.4	Splice site loss
			compound HET	*CDK5RAP2*	chr9:123205949	**0.01%**	p.V1033fs*41	c.3097delG	NM_018249.4	frameshift
BRC_90	Microcephaly and a simplified cortex	F	compound HET	*ASPM*	chr1:197057461	0.00%	p.D3363fs*18	c.10086delA	NM_018136.4	frameshift
			compound HET	*ASPM*	chr1:197102623	0.00%	p.R759delinsL*VKQTTW	c.2275_2276insTGTAGGTTAAACAGACTACGT	NM_018136.4	frameshift
Potentially clinically relevant findings										
BRC41	Microcephaly, seizures and hyperphosphatasia (consanguineous).	M	homozygous	*PGAP3*	chr17:37842179	0.00%	p.G92D	c.275G>A		Missense
BRC_146	Microcephaly, trigonocephaly and cerebellar vermis atrophy.	M	*de novo*	*SOX11*	chr2:5832940	0.00%	p.C29*	c.87C>A	NM_003108.3	Stop gain
BRC_231, 1139F1, 1139E1	Three fetal samples from the same family with polymicrogyria, cerebellar hypoplasia and bilateral talipes	F, F, M	compound HET	*PI4KA*	chr22:21067580	0.00%	p.D1854N	c.5560G>A	NM_058004.3	Missense
			compound HET	*PI4KA*	chr22:21119924	0.00%	p.Arg796Ter	c.2386C>T	NM_058004.3	stop_gained

Table 3.15: continued from previous page

Patient ID	Patient / Family Phenotype Summary	Gender	Established inheritance pattern	Gene	Coordinates	EVS freq	Protein Variant	Transcript Variant	Transcript ID	Translation Impact
Candidates but require further functional studies or patients with similar variants										
BRC_224	Microcephaly and seizures (consanguineous).	F	homozygous	*PSMB1*	chr6:170855244	0.00%	p.G84S	c.250G>A	NM_002793.3	Missense
BRC_225	Lissencephaly with contractures and microcephaly.	F	*de novo*	*CHMP7*	chr8:23117757	0.00%	p.T408fs*6	c.1222dupA	NM_152272.3	frameshift
BRC_230	Severe microcephaly (-5 SD) (consanguineous).	F	homozygous	*GPT2*	chr16:46934600	0.00%	p.L115fs*67	c.340_346delGCACTAT	NM_133443.2	frameshift
BRC_226	Microcephaly, seizures and hypocalcaemia (consanguineous).	M	homozygous	*CLIP2*	chr7:73790701	0.00%	p.K622T	c.1865A>C	NM_003388.4	Missense
BRC_91	Microcephaly, short stature, and learning difficulties.	M	*de novo*	*PHF7*	chr3:52457227	0.00%	p.P347L	c.1040C>T	NM_016483.5	Missense
BRC102	Microcephaly (consanguineous)	M	homozygous	*PIGY*	chr4:89444948	0.00%	5' UTR promoter region	c.-540G>A	NM_001042616.2	Unknown (promoter region)
BRC180	Microcephaly (consanguineous)	M	homozygous	*CLASP1*	chr2:122125392	0.00%	p.G1161S;	c.3457G>A	NM_001142273.1	Missense
BRC50	Microcephaly and cerebellar atrophy (consanguineous).	F	homozygous	*ACADSB*	chr10:124802613	0.00%	p.H245Y	c.733C>T	NM_001609.3	Missense
Families with no candidate variants										
BRC_83, BRC_84	Two siblings with microcephaly.	M,M	unknown							
BRC_228	Microcephaly, seizures and raised ALP.	M	unknown							
Family with previous clinical diagnosis										
BRC_232	Child with 22q11 and polymicrogyria.	M	*de novo* occurrence of CNV							
Sequencing Failed										
BRC_153	Angelman syndrome like, microcephaly, seizures.		unknown							

3.2.2. Clinically relevant brain abnormality findings for patients and their families

In this project variants were identified in known brain abnormality genes. These were assessed as clinically relevant based on or more of the following criteria: i) the patient's phenotype closely matched that of previously described patients; ii) the variant was previously reported as pathogenic or in the same exon as previously described pathogenic variants; iii) the predicted effect of the variant on protein coding was severe; iv) the variant was at very low frequency or not present in common variation databases. These criteria were based on in house experience for identifying pathogenic variation and adhered to the ACMG recommendations for interpretation and reporting of sequence variation [157].

3.2.2.1. *ASPM*

BRC_90 was not reported as consanguineous and a trio-based approach was undertaken to exome sequence the proband and parents. BRC_90 presented with microcephaly and a simplified cortex, which matches the phenotype of patients reported with pathogenic *ASPM* variants. Consistent with this, compound heterozygous variants were identified in *ASPM*. The variants were both frameshifts predicted to severely effect protein coding. Segregation in the family was consistent with compound heterozygosity and was validated with Sanger-sequencing. An inclusion criterion of this project was that the patients were pre-screened for variants in *ASPM,* as it is the most common cause of autosomal recessive microcephaly [74]. This result suggests the variants were missed by prior Sanger sequencing. In the WGS500, there was a 2% false negative rate of genes previously screened by traditional Sanger sequencing (manuscript in preparation). This

was further evidence that an NGS approach can be more sensitive than Sanger sequencing.

Since the predicted effect of the variants on protein coding were severe and both were in exons in which pathogenic variants had previously been described either in HGMD or ClinVar, they were classified as causative. The data has been sent to the RGL where the variants will be validated by Sanger sequencing to provide a clinical diagnosis for the family. There is an affected sibling in this family with a similar phenotype. The DNA was unavailable for this project, but future clinical testing in the RGL will ascertain if the sibling inherited both *ASPM* variants.

3.2.2.2. *CASK*

In this cohort, two unrelated female patients were identified with *de novo* frameshift variants in *CASK*, which encodes the calcium dependent serine protein kinase. Both patients were exome sequenced along with their parents and Sanger sequencing confirmed the *de* novo origin of the variant in the families. Mutations in *CASK* lead to a condition which belongs to a group of X-linked disorders thought to result in reduced male viability or *in utero* lethality [158]. Heterozygous *CASK* mutations are associated with the phenotype of mental retardation and microcephaly with pontine and cerebellar hypoplasia [159]. The phenotypes of BRC_16 and BRC_59 match the phenotypic spectrum associated with *CASK* variants [159]. Both variants were in exon 17 of *CASK*, which could be coincidental or it is possible that this region is a frameshift mutation hotspot which has been observed in other disorders [160]. The coding sequence surrounding each variant was examined but nothing was found to explain the potential

hotspot. There was an additional pathogenic variant in exon 17 reported in ClinVar [161], which suggests that the exon is important for proper protein function. Both of the variants were predicted to have severe effects on protein coding and based on the above evidence, were classified as causative. The information has been sent to the RGL for the variants to be clinically validated in the two probands.

3.2.2.3. *CDK5RAP2*

Since consanguinity was not reported, a trio based approach was undertaken to exome sequence BRC_81 and both parents. Compound heterozygous variants were identified in BRC_81 in *CDK5RAP2* [162], another known microcephaly gene. The variants and segregation were validated with Sanger-sequencing. The patient presented with microcephaly and cortical dysplasia, which is similar to the phenotype of previously reported *CDK5RAP2* patients [82,162,163]. The first variant was a single base deletion predicted to cause a frameshift and premature stop codon in exon 23. The variant was at 0.01% frequency in NHLBI ESP database of European individual exomes. The second was a splice site variant which was a potential splice donor loss. An RNA sample has been requested for the patient to ascertain if the variant affects splicing. Since it is a known gene that closely matches the patient's phenotype, the variant was classified as likely to be causative and reported to the RGL to be validated. Since *CDK5RAP2* is a known microcephaly gene it does raise the question if it should have been pre-screened for this study. The NHS currently only has clinical tests for variants in *ASPM* and *WDR62* as these are the most common causes of autosomal recessive microcephaly. There are only a small number of patients reported in the literature with *CDK5RAP2* variants and the initial patients to be described in the literature were all

Pakistani[162,163]. Since the family is Caucasian, originating from Eastern Europe, the referring clinician did not suspect *CDK5RAP2* variants. It does highlight that the prevalence and distribution of *CDK5RAP2* variants might be higher than previously thought and the importance of not introducing bias in candidate gene selection.

3.2.2.4. *PCDH19*

BRC_115 was referred to Clinical Genetics presenting with microcephaly, epilepsy and significant developmental delay. The patient was recruited into the brain abnormality study based on the microcephaly phenotype. Since there was no reported family history of the phenotype or consanguinity, a trio-based approach was undertaken to sequence the proband and parents. A variant was identified in *PCDH19* in BRC_115 and Sanger sequencing confirmed that the variant had arisen *de novo* in the proband. *PCDH19* is implicated in Epilepsy Limited to Females with Intellectual disabilities (EFMR) [164]. The epilepsy and developmental delay match the phenotypic spectrum associated with *PCDH19* loss of function variants. Microcephaly is not normally characteristic of EFMR but there is a clinical report of a patient presenting with post-natal microcephaly (-2SD) [165]. The microcephaly phenotype of BRC_115 was examined with the clinician and it was revealed that the patient has 'relative' microcephaly. The patient's head size is at the 25th centile but their height and weight is in the 75th centile. The clinical definition of microcephaly is widely interpreted, but it is normally a reduced occipitofrontal circumference (OFC, head circumference) of at least -2SDs but usually -4SDs from the normal [74]. Since the patient does not have true microcephaly, their phenotype fits in the spectrum of *PCDH19*.

The particular variant in this case was predicted to cause the protein change, p.N340S which has been reported in multiple cases of EFMR [164,166,167]. The variant sequence was examined to assess whether it could be a mutation hotspot, which would explain the multiple instances of the variant arising *de novo*. The sequence surrounding the variant, TTGTCATTGG, was repeated nearby in the same exon (codons 231-234). Short repetitive DNA elements have been implicated in mutation hotspots [168]. It is possible that this short element is responsible for the repeated origin of the variant or there could be another mechanism at play. The variant in BRC_115 was classified as pathogenic based on it being a known causative variant and reported to the RGL for clinical validation.

3.2.3. Variants in Glycosylphophatidylinositol (GPI) pathway gene *PGAP3* implicated in brain abnormality phenotype for BRC_41 and siblings

The result exemplar of this thesis was the identification of a homozygous deleterious variant in *PGAP3* affecting the penultimate stage of GPI biosynthesis. BRC_41's family was recruited into the research study based on the brain abnormality phenotypes they exhibited. The family originates from the Sargodha district of Punjab in Pakistan. All three children (V-1, V-2 and V-3 in Figure 3.10) presented with profound developmental delay, severe learning disability, no speech, psychomotor delay and postnatal microcephaly (-2 to -3SDs). The two older children developed generalized tonic clonic seizures at ages 12 and 4 years.

3.2.3.1. Experimental approach for identification of candidate variants

An autozygosity mapping approach was undertaken as consanguinity was reported in the family and the DNAs from the entire nuclear family were therefore run on CytoSNP-12 arrays. The parent-parent-child (PPC) heritability ratios were calculated in Genome Studio (Illumina), which assesses the inheritance of each individual SNP for Mendelian errors and confirmed the familial relationships.

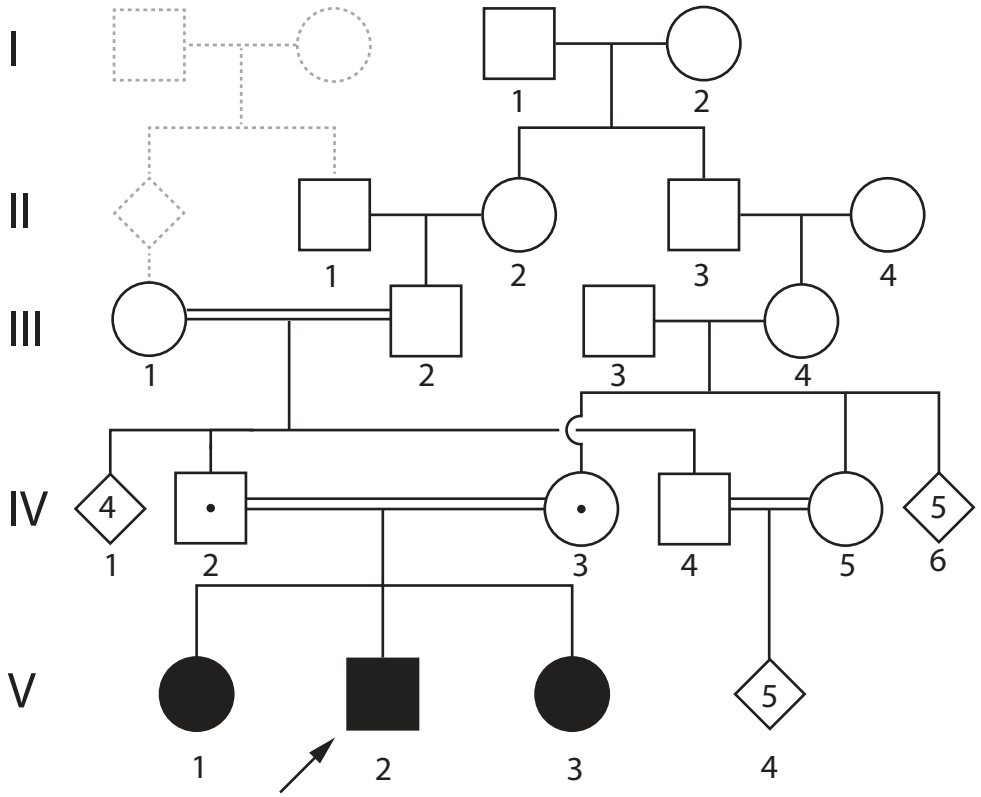


Figure 3.10: Pedigree of BRC_41
Pedigree is shown for BRC_41 (V-2) and family who are of family of Pakistani descent. Shading indicates intellectual disability and hyperphosphatasia and proven heterozygote carriers are shown by a dot. Grey dotted lines indicates an additional consanguineous loop not described in the clinical notes, but which was inferred by the high inbreeding coefficient detected in IV-2 from the SNP array data. The proband BRC_41 is indicated with an arrow.

Chapter 3: Results

The SNP array data was analyzed in Nexus, as described in the Methods section 2.4.1. There were no CNVs of likely clinical relevance, which confirmed the result from the BRC's previous unpublished aCGH analysis. The inbreeding coefficients were estimated based on the total regions of homozygosity called by Nexus (Table 3.2). It was originally thought that the parents were first cousins, but the inbreeding coefficients implied that they were second cousins and this was confirmed with the extended family pedigree (Figure 3.10). Interestingly, the coefficient of inbreeding in the father suggested that the paternal grandparents were first cousins (see grey Figure 3.10), which was not confirmed by the referring clinician. The shared regions of homozygosity (ROH) in the siblings were inspected visually in Nexus. It was evident that all three affected children only shared one region of homozygosity larger than 0.5Mb (Figure 3.11). The specific shared regions were determined with HomozygosityMapper and PLINK as described in the Methods section 2.4.2. The large ROH was 13.8Mb, from 31,385,743 to 45,184,794 on chromosome 17. The borders of the region were accepted as the larger of estimates of each from PLINK and HomozygosityMapper. A recombination event in BRC42 defined the distal end of the homozygous region (Figure 3.11). There were no striking microcephaly or brain abnormality candidates identified among the 374 refseq-annotated genes in the region so a next generation sequencing approach was undertaken to WES one of the siblings in the family. The quality of the exome sequencing was assessed as described in the Methods and found to be of high quality based on the high genotype concordance with SNP array data (see Section 3.1.5).

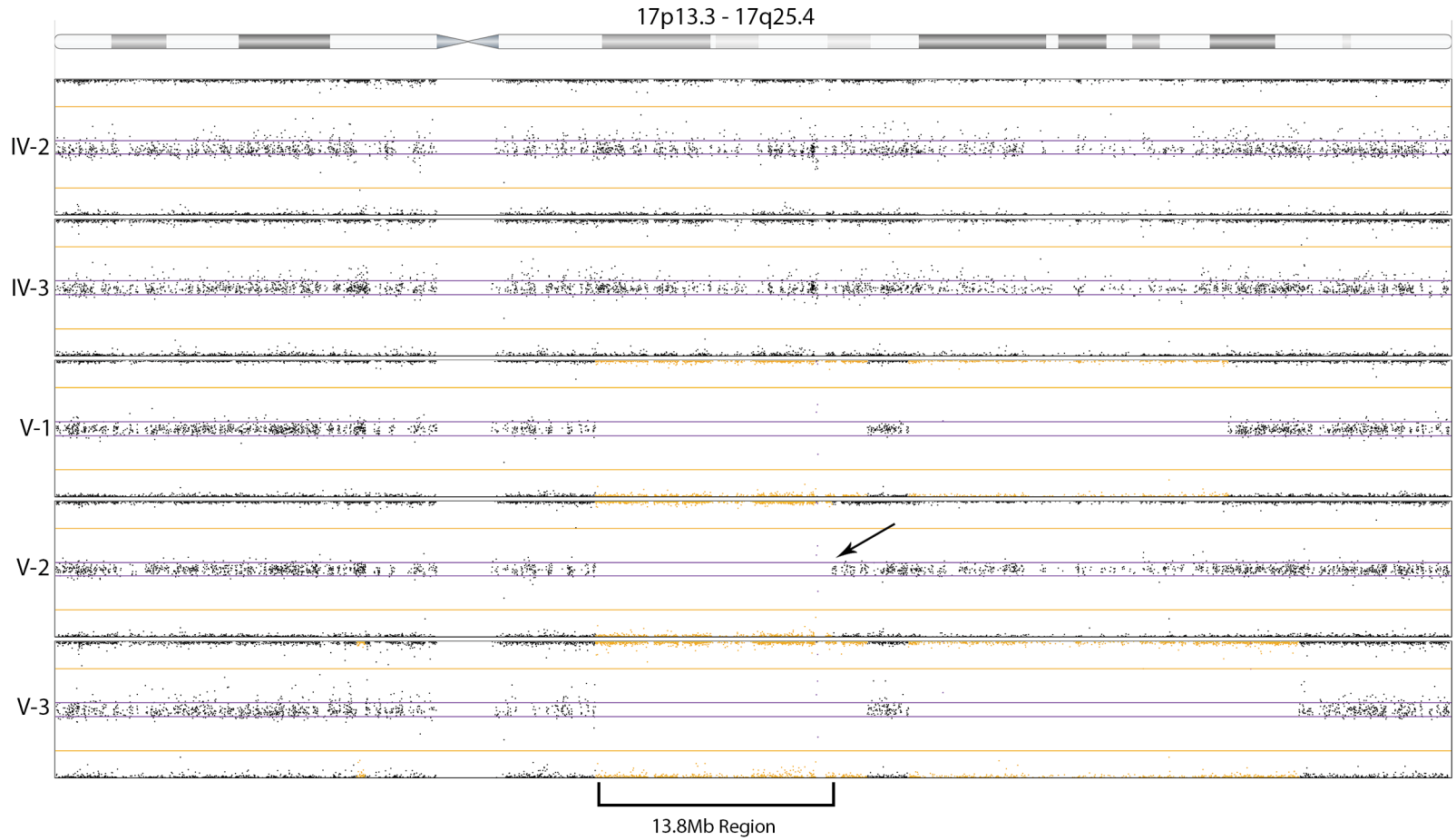


Figure 3.11: Shared region of homozygosity in BRC_41 and affected siblings.
Analysis of B allele frequencies for chromosome 17 shows a 13.8Mb region of autozygosity shared by all three affected children in BRC_41 family. Based on the GRCh37/hg19 reference sequence, the genomic coordinates of this candidate region were chr17:31,385,743 – 45,184,794. Image created using Nexus Copy Number v7.0. Arrow indicates recombination event that defines the distal end of candidate region. Homozygous genomic segments were identified using Nexus Copy Number software and compared across the three affected individuals. A 4 Mb minimum cutoff as regions measuring up to that size are common in demonstrably outbred individuals. Results were confirmed by running the SNP array data through the HomozygosityMapper tool and PLINK.

3.2.3.2. Identification of *PGAP3* variants from the analysis of Whole Exome Sequencing data of BRC_41

The VCF for BRC_41 was created with Platypus and uploaded to IVA and was analyzed as described in Methods section 2.7.5. It was filtered in IVA for any rare, deleterious variants in the shared ROH (Figure 3.12). The large ROH on chromosome 17 was found to harbour a rare missense variant in *PGAP3.* The variant c.275G>A, predicts a p.G92D alteration affecting a highly conserved position in the protein sequence (Figure 3.13). The variant was validated with Sanger sequencing and segregated in the family (Figure 3.14). This provided further evidence that the candidate region was homozygous by descent, as predicted by the SNP arrays.

Variants	Genes	
199,971	17,540	
150,877	16,954	Call quality >40 and read depth ≥8x
12,966	3,764	< 1% frequency and not in dbSNP 138
179	160	Predicted Deleterious
7	7	Homozygous
1	1	17q11.2-q21.32 candidate region

Figure 3.12: Ingenuity variant filtering cascade for BRC_41
Ingenuity Variant Analysis Filtering cascade for Exome Sequencing.

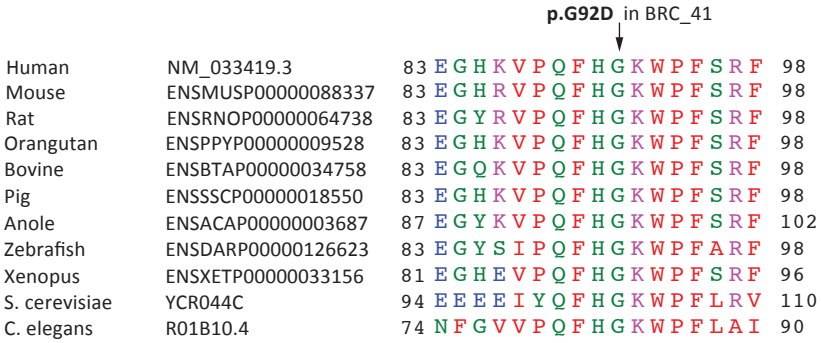


Figure 3.13: Phylogenetic Conservation of *PGAP3* variant p.G92D
Clustal W alignment of amino acid sequence shows conservation at p.Gly92 down to C. elegans.

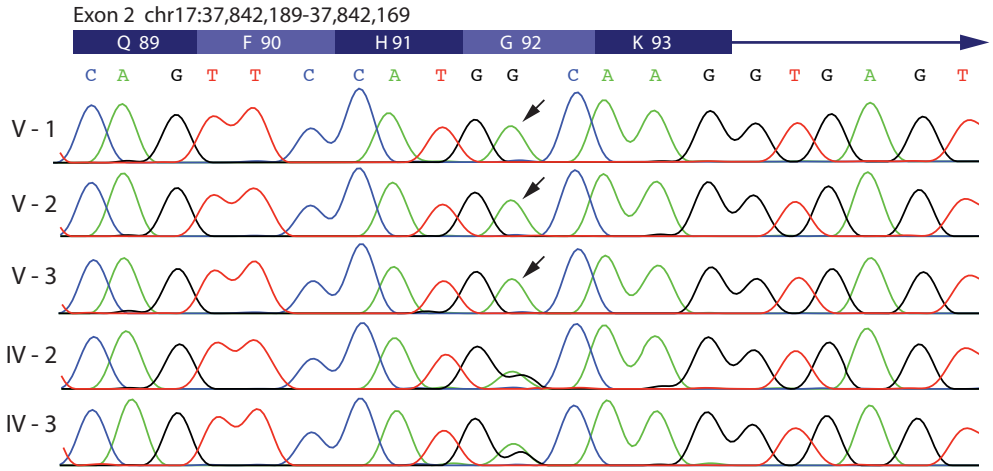


Figure 3.14: Sanger validation of *PGAP3* variants in BRC_41 and family
Sanger validation and segregation testing of *PGAP3* variant. Proband BRC_41 is labelled as individual V - 2. . DNA for the unaffected double first-cousins (see Figure 3.10, collectively labelled as V - 4) was not available.

The variant was not present in the NHLBI ESP Exomes, or 1000 Genomes datasets, or Public Complete Genomics Genomes, or 274 in house genomes of mixed ancestry (WGS500). These databases include only a few known individuals of Pakistani descent. It was possible that the variant was a polymorphism specific to the Pakistani ethnic group, which if at high frequency could imply that it was benign. In order to test this, the recently released data from Phase III of the 1000 Genomes Project was downloaded for 108 Punjabi individuals from Lahore, Pakistan. I ran Platypus on the 108 low coverage BAM files with default settings. The variant was not called in any of the samples. Since the average coverage was low in the samples, 8.5x at the variant locus, it was possible that Platypus might have missed heterozygous variants. The BAM files were also inspected visually and two samples were found to contain the variant in a single read with low mapping quality and PHRED scores. The DNA for these samples was obtained from the Coriell Cell Repository and Sanger sequencing confirmed that the variant was not present. Based on this data it is unlikely that c.275G>A is a population specific polymorphism.

The variant was predicted to be deleterious by SIFT (0.000), PolyPhen-2 (0.998) and to be disease causing by MutationTaster (0.99) [169]. To strengthen the evidence of the variant being pathogenic in this family, the rest of the exome data was analyzed for rare deleterious mutations. No variants were found directly associated with brain abnormalities or seizures, based on both recessive and dominant inheritance models. There were other rare variants, which were all carefully reviewed. They were discarded as candidates based on either i) lack of association with the phenotype ii) the variant

being sequencing noise when inspected in IGV iii) the variant was not in an identical region by descent among the three siblings from the array data. It was clear that the variant in *PGAP3* was the strongest candidate based on autozygosity mapping and its absence in all of the genetic variation databases. *PGAP3* encodes a trans membrane protein involved in the fatty acid remodelling the GPI anchor in the Golgi apparatus. It is part of the GPI pathway as described below.

3.2.3.3. Overview of GPI Pathway and previously reported GPI disorders

Glycosylphophatidylinositol (GPI) anchoring is a post-translational modification that tethers proteins to plasma membranes, and is thought to play a role in protein sorting and trafficking [170]. The GPI anchor is well conserved among eukaryotes and there are over 150 mammalian GPI-anchored proteins (GPI-APs), including receptors, adhesion molecules, and enzymes [170]. Many of these proteins are critical for normal neural and embryonic development [171,172].

There are around 30 genes known to be involved in the biosynthesis and remodelling of the GPI anchor. The GPI-anchor is formed in the endoplasmic reticulum (ER) and attached to a protein showing a specific C-terminal signal before it is transported to the Golgi apparatus for fatty acid remodelling and cellular export. A defect in the C-terminal signal can result in a protein not being properly GPI anchored and as a result not transported to the correct tissue [173]. It is likely that as the GPI pathway becomes more widely understood more defects of that nature will be identified.

Chapter 3: Results

In the last four years germline mutations in 11 genes of the GPI anchor synthesis pathway have been shown to cause disorders with a wide spectrum of phenotypes including seizures, congenital abnormalities, brain malformations and intellectual disability. A number of disease causing mutations in GPI-pathway genes were first identified using traditional autozygosity mapping in combination with the next generation sequencing methodologies [174,175]. All individuals with a congenital impairment of the GPI-anchor synthesis have residual surface levels of GPI-anchored proteins, and it is hypothesized based on animal models that complete loss of function mutations in genes of the GPI-anchor synthesis are embryonically lethal [176].

Alkaline Phosphatase (ALP) is a GPI-AP that is used as a metabolic marker for GPI deficiency disorders. GPI disorders can be subdivided into groups based on their effect on serum levels of ALP. Mutations that affect the early GPI pathway genes inhibit the formation of the GPI molecule and the attachment of proteins. These variants do not result in high ALP due to the intracellular degradation of the growing GPI molecule and protein [177]. The mutations that affect the later stage GPI pathway result in increased serum ALP due to proteins being cleaved from the growing GPI molecule and released into the cytoplasm by GPI transamidase [177].

Recently, mutations in *PGAP2* have been reported that affect the final GPI anchor fatty acid remodelling step resulting in abnormal GPI-APs [174,178,179]. GPI remodelling occurs in the Golgi and involves the removal of an unsaturated fatty acid at the sn-2 position of the GPI anchor by *PGAP3* followed by the addition in its place of a saturated

fatty acid by *PGAP2*. The fatty acid remodelling is critical for proper association of GPI-APs with lipid rafts [180]. The reports of pathogenic variants in *PGAP1* and *PGAP2* in patients with severe disabilities suggest that the final stages of GPI-anchor biosynthesis involving *PGAP3* (which acts immediately upstream of PGAP2) might also be necessary for normal human development.

3.2.3.4. Evidence for Pathogenicity of *PGAP3* variants

During the research on this family, two concurrent studies were published showing that *PGAP2* was associated with Hyperphosphatasia with Mental Retardation Syndrome [174,178]. Since patients with *PGAP2* variants presented with hyperphosphatasia, and it comes after *PGAP3* in the GPI pathway, serum alkaline phosphatase (ALP) testing was therefore sought for the affected children in this family. The exact mechanism of release ALP differs among the GPI pathway disorders. In brief, ALP is a GPI-anchored protein. When the anchor is not properly formed in the later steps of GPI maturation, ALP is released into the serum instead of being anchored on the lipid cell surface. The increased concentration of ALP in the blood can be detected with a standard metabolic blood test. In this project, blood from the whole family was tested in an OUH Biochemistry Laboratory. The ALP levels were tested several times and found to be elevated above the normal in all three children. The values ranged from 739 U/l to 1452 U/I, with the normal upper limit of the ALP test specified as 525-600 U/l. Raised serum ALP is an indicator of a GPI deficiency but it is not a conclusive test. There are other potential reasons for isolated high ALP measurements, such as growth spurts [181]. Further evidence was required to assess the implication of the *PGAP3* variants on brain abnormalities and human disease.

To confirm the effect of the *PGAP3* variant on GPI synthesis, collaboration was sought with Professor Taroh Kinoshita based at Osaka University. The BRC research group at Oxford had previously collaborated with Professor Kinoshita's research group in functional work on the *PIGQ* variant from the WGS500 EIEE study [94]. The Kinoshita group had also published the initial papers describing several of the GPI genes and specialized in functional testing and analysis of variants in GPI pathway genes. Several prediction algorithms were used to model the effect of the p.G92D variant on protein structure [182]. This data was difficult to interpret without in depth knowledge of *PGAP3*'s structure and expertise in the GPI pathway. It was hypothesized by Professor Kinoshita that the variant introduces a negatively charged amino acid into the conserved region adjacent to the lipid membrane. The negative charge could potentially repel the growing GPI molecule and cause it to miss the fatty acid remodelling step of the pathway (Figure 3.15). It had already been shown with *PGAP2* variants that this step was critical in the GPI pathway[174,178]. To confirm this hypothesis, functional work was undertaken by the Japanese group to determine the consequence of the *PGAP3* variants on the GPI pathway.

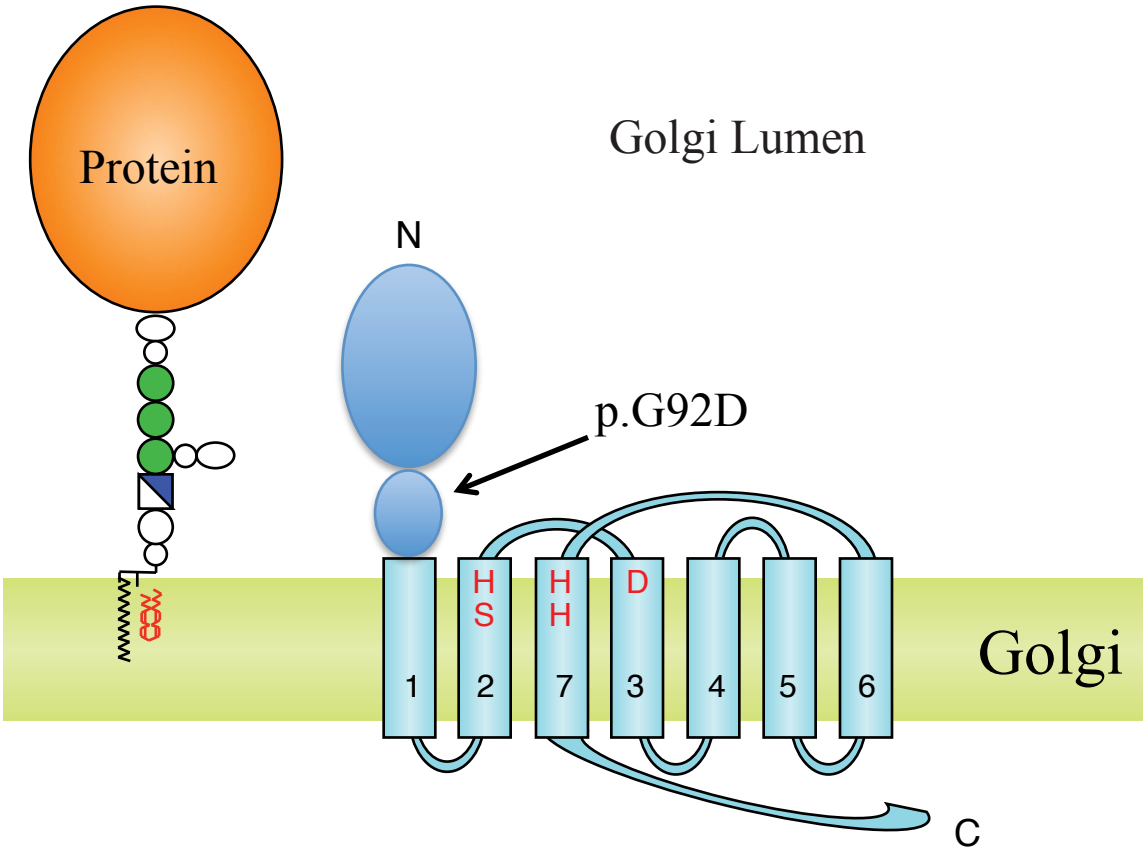


Figure 3.15: Representation of growing GPI molecule and approximate position of mutation in *PGAP3*

p.G92D variant could potentially repel the growing GPI molecule since it introduces a negatively charged amino acid into the conserved region adjacent to the lipid membrane. Reproduced with permission from Taroh Kinoshita.

Chapter 3: Results

A knockout *PGAP2* and *PGAP3* Chinese hamster ovary (CHO) cell model had been established as part of the Kinoshita group's original work defining the fatty acid remodelling step of the GPI pathway [180]. The effect of the *PGAP3* variants on the GPI pathway was assessed by the Japanese group using this model. The principle of this assay is that defects in the GPI pathway will result in lowered cell surface expression of GPI-anchored proteins which can be measured using flow cytometry. To assay the effect of the *PGAP3* variants, the deficient CHO cell lines were transfected with wild type or variant *PGAP3*. The cells were subsequently stained with fluorescent-labelled antibodies that targeted GPI anchored proteins (GPI APs). The relative florescence of the molecules was assessed using a flow cytometer. In this model when cells are transfected with wild type *PGAP3* they show decreased cell surface expression of GPI-anchored proteins. When the cells are transfected with defective *PGAP3*, the cell surface expression is only mildly reduced compared with the double knockout. The model was designed in this manner as GPI molecules that miss the PGAP3 remodelling step still have two fatty acid anchors and are not immediately released from the cell surface [180]. The downstream mechanism that causes GPI molecules lacking PGAP3 mediated remodelling to be released from the cell surface has yet to be characterized [180], but it does occur based on the hyperphosphatasia phenotype of conditional *PGAP3* knockout experiments in mice. In a normal cell, a functioning *PGAP3* protein removes one of the unsaturated fatty acid anchors and *PGAP2* replaces it with a saturated fatty acid. By knocking out *PGAP2*, those molecules that undergo removal of one fatty acid by functioning *PGAP3* are immediately released due to cleavage by phospholipase activity [180]. Since the GPI APs are immediately released, the change in cell surface expression can be measured by the

flow cytometer. Since a completely defective protein would show no change in cell surface expression, the transfection efficiency was assessed by real time PCR expression of the defective *PGAP3*.

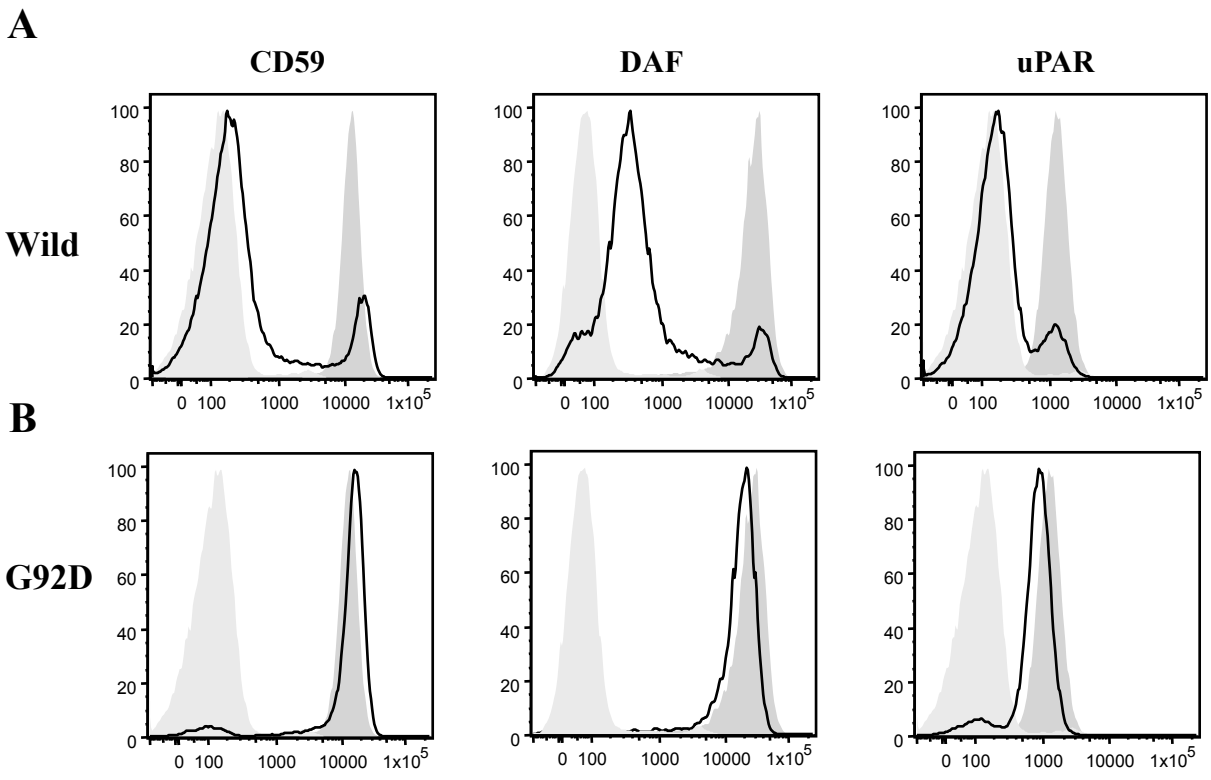


Figure 3.16: Flow cytometric analysis for function of *PGAP3* variant in BRC_41
The graphs show the results of the knockout *PGAP2* and *PGAP3* Chinese hamster ovary (CHO) cell assay for the function of p.G92D (all of the functional work was done by Professor Kinoshita's group in Japan). CD59, DAF(CD55) and uPAR are GPI anchored proteins. CHO cells were stained with fluorescent-labelled antibodies for CD59, DAF and uPAR to assess the cell surface expression of GPI AP.
(A) After transfection with wild-type *PGAP3* cDNA, the surface levels of three GPI-APs were severely reduced (solid lines). The dark area shows the original surface levels of GPI-APs on *PGAP3* and *PGAP3* double mutant CHO cells, whereas the light grey area represents the isotype-matched control.
(B) PGAP3 cDNA bearing mutation c.275G>A (p.G92D) from BRC_41 did not reduce or only slightly reduced API-AP levels (solid lines) indicating that the variant causes a null or nearly null phenotype.

The functional work by the Kinoshita group in Japan using the model explained above strongly supported the pathogenicity of the p.G92D PGAP3 variant (Figure 3.16). The variant did not reduce or only slightly reduced the cell surface expression of the GPI APs, which implies that the variant has the same effect as completely deficient *PGAP3*. The functional evidence showed that the *PGAP3* variants in BRC_41 and siblings were likely to cause a GPI deficiency disorder. Many of the phenotypic features of the three affected children were consistent with *PGAP2* patients and previously reported variants in patients with deficiencies in other GPI-pathway genes [174,178]. Based on the functional work, the similarity of phenotypes to other GPI deficiencies and the novelty of the variant, *PGAP3* was the likely causative agent of the brain abnormality phenotypes.

In an attempt to identify further patients with *PGAP3* variants, all 8 exons of *PGAP3* were Sanger sequenced in a cohort of 12 patients with previously unexplained developmental delay and seizures. Of these, BRC_228 was recruited for this project specifically due to a *PGAP3* like phenotype. No *PGAP3* variants were identified in any of the patients. A collaboration was then formed with Dr Peter Krawitz in Germany who had developed a panel to test GPI genes in patients presenting with hyperphosphatasia and developmental delay. The pathogenicity of the *PGAP3* variants was strengthened by Krawitz identifying two unrelated individuals, one of consanguineous union with homozygous variants, and one with compound heterozygous variants in *PGAP3*. The phenotypes of all the affected individuals was compared and the common features noted were severe psychomotor delay with no speech, marked motor delay as well as

characteristic facial features **(**Figure 3.17**)** and normal growth parameters and OFC at birth. All of the individuals had consistently elevated ALP which varied between 1.4 to 5 times of the age adjusted upper limit of the normal range. Four of the 5 individuals developed seizures at ages ranging from 18 months to 12 years.

Thus, all of this evidence taken together suggests that the candidate variant in *PGAP3* is the cause of the brain abnormality phenotype. This research on *PGAP3* has been published in the February 2014 issue of American Journal of Human Genetics [6].



Figure 3.17: Photographs of *PGAP3* affected individuals.
Facial features seen in (A) individual V-3 from BRC_41 family and individuals in (B) and (C) bear a striking resemblance, with broad nasal bridge, upslanting palpebral fissures, broad nasal tip, short nose, long philtrum, thin and wide upper lip, full cheeks and large fleshy ears, and soft skin. Images are shown with consent. Although facial features of individuals V-1 and V-2 (BRC_41) are similar to those seen in their younger sister, consent was not given to show images from these individuals.

3.2.4. Identification of *PI4KA* variants in polymicrogyria family (1139E1, 1139F1, BRC_231)

An important result of this thesis was the identification of a potential novel disease gene in a single family with three terminated fetuses (Figure 3.18). Since consanguinity was not suspected, the entire nuclear family was exome sequenced. Compound heterozygous variants were identified in all three fetuses in an important phosphatidylinositol kinase, *PI4KA*. The fetuses were diagnosed in utero with severe congenital malformations and the pregnancies were terminated. The post-mortem pathologist report highlighted that that the fetuses had polymicrogryia (PMG), a neuronal migration disorder resulting in the abnormal development of the brain's folds, gyri. The disorder is characterized by an excessive number of smaller than normal gyri. The specific details of the phenotypes of 1139E1 and 1139F1 were PMG, cerebellar hypoplasia, severe bilateral talipes and small lungs. One of the probands was also large in size, 98th centile and had overlapping fingers. The third fetus, BRC_231 had PMG and presented with a similar phenotype but there has yet to be in depth phenotypic report from the clinician.

3.2.4.1. Approach for candidate disease gene identification

Since the family had not been tested for CNVs, one of the fetuses, 1139E1 and both parents were screened on cytoSNP-12 arrays. Only 1 of the fetuses was screened due to limited availability of DNA. There was one small duplication detected on chrX:6,467,204-6,838,627 which was maternally inherited and did not contain any genes. An X-linked disorder was not suspected, as there were male and female fetuses. Therefore this duplication was not considered pathogenic and all three siblings and parents were exome sequenced.

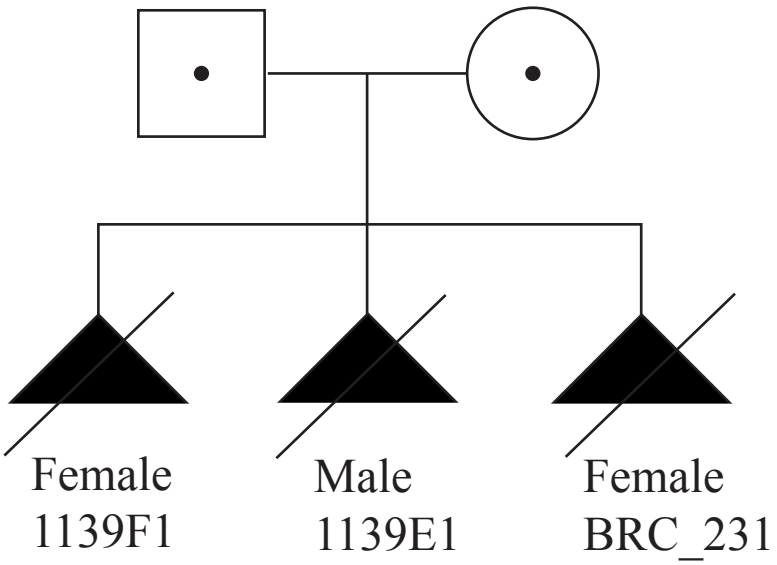


Figure 3.18: Pedigree of PMG family
Pedigree is shown for 1139F1, 1139E1, BRC_231 and parents (BRC_219 and BRC_220). Shading indicates termination of pregnancy due to severe congenital abnormalities (gestational times unknown). Proven heterozygote carriers of *PI4KA* compound heterozygous variants are shown by a dot (father c.2386C>T and mother c.5560G>A). WES was performed on all members of the family.

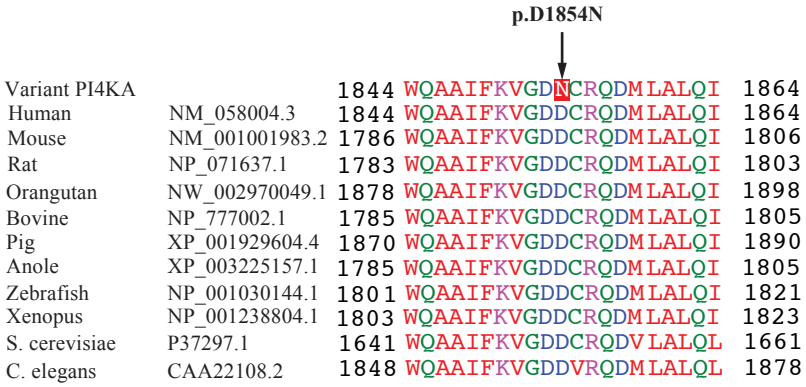


Figure 3.19: Phylogenetic Conservation of *PI4KA* variant p.D1854N
Clustal W alignment of amino acid sequence shows conservation at p.D1854 down to C. elegans.

The WES data from the family was extensively examined in IVA for any shared variants among the three fetuses. Variants in *PI4KA* were the only rare deleterious variants shared by all three probands that fitted with the presumed autosomal recessive inheritance pattern. Neither variant was reported in the NHLBI ESP Exomes, or 1000 Genomes datasets, or Public Complete Genomics Genomes, or 274 in house genomes of mixed ancestry (WGS500). The first variant in *PI4KA* was predicted to cause a nonsense mutation in exon 20 and the second variant, a missense change p.D1854N was in exon 48 in the important kinase-binding domain of the protein [183]. ClustalW alignment of amino acid sequence showed conservation at p.D1854 down to C. elegans (Figure 3.19). MutationTaster (0.99), SIFT (0.003), and PolyPhen-2 (1.000) predicted the variant to be disease causing and damaging. The stop gain variant validated by Sanger sequencing in all three fetuses but the validation of the missense variant was complicated by the presence of two non processed pseudo-gene copies of *PI4KA, PI4KAP1* and *PI4KAP2* less than 1Mb upstream and downstream. These pseudo-genes appear to only have originated from a selection of exons. *PI4KAP2* includes surrounding intronic sequences and the exons 33-34-35-36, 44-45 and 48-49-50-51. *PI4KAP1* is similar but also includes exons 52-53-54-55. It was difficult to validate the missense variant, as exon 48 was present in both pseudo-gene copies. The Sanger sequencing was relatively clear but there appeared to be some signal from the pseudo-gene copies (Figure 3.20).

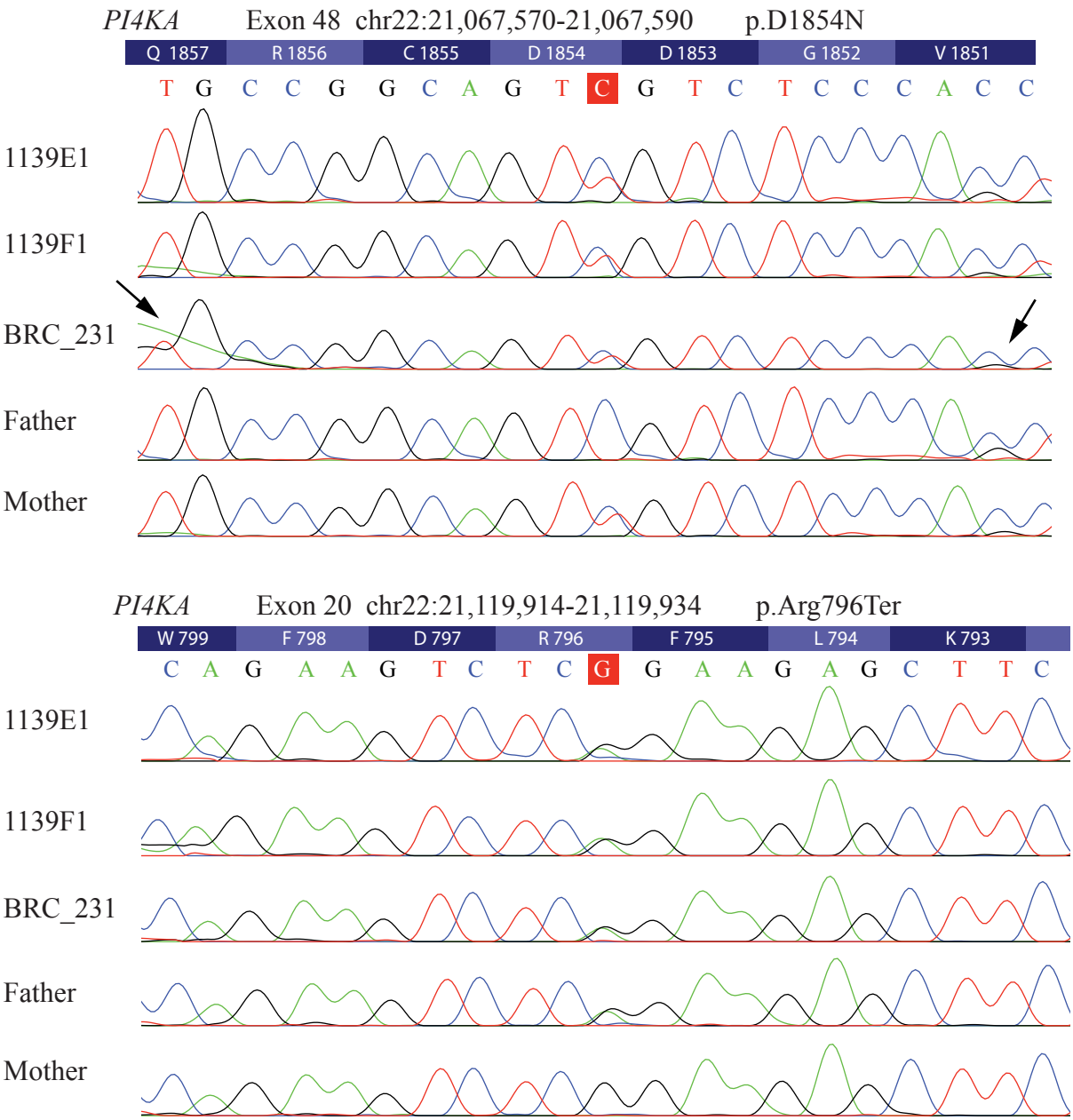


Figure 3.20: Sanger validation of *PI4KA* mutations
Sanger validation and segregation testing of *PI4KA* variants. Arrows indicate noise in sequencing of the missense variants.

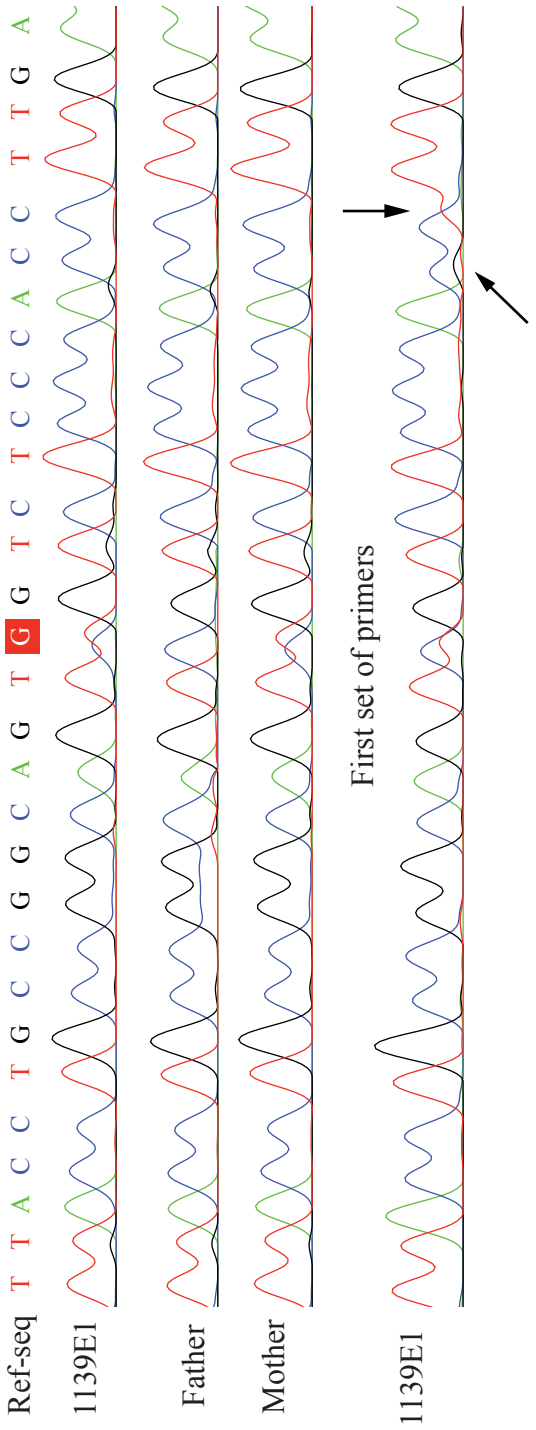


Figure 3.21: Comparison of Sanger traces of *PI4KA* variant p.D1854N with new primer
Traces appear slightly cleaner with re-designed primer, arrows indicate regions of noise in original set of chromatograms that were suspected to be due to pseudo genes *PI4KAP1* and *PI4KAP2*.

It was a concern that the variants detected by the exome sequencing could actually be a result of the pseudo-genes. The Sanger sequencing primers were re-designed to bind to a region where *PI4KA* differed from the pseudo-gene copies. The subsequent Sanger sequencing was cleaner, but ultimately still slightly noisy (**Figure 3.21** and Figure 3.22).

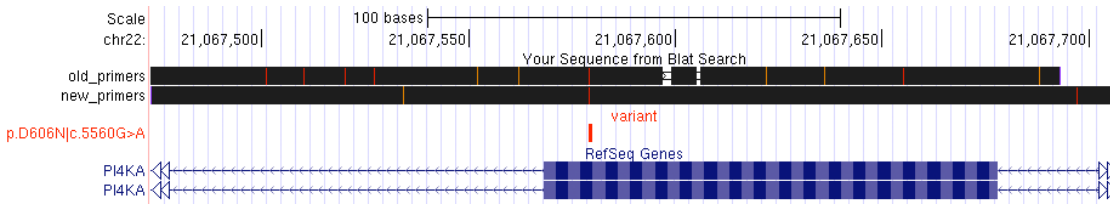


Figure 3.22: Comparison of BLAT results from sequences using old and re-designed primers.
Re-designed Sanger primer shows significantly improved alignment with *PI4KA* missense variant region. The consensus sequences obtained from Primer4 were run through the UCSC BLAT tool. The black bars represent the aligned sequences and the red lines and gaps in the bars are regions that do not match the reference sequence.

In another approach, the forward and reverse exome sequencing read pairs from the variant region were aligned to the reference using the UCSC BLAT tool [95]. It was possible to map the read pairs to the pseudo-gene regions with several errors (Figure 3.23), but it was clear that the reads mapped better in the actual gene. It was also likely that if the missense variant was noise from the pseudo-genes that it would have been observed in another of the 48 exomes from this project, in the WGS500 study or in the NHLBI exome variant database. It was concluded that the variant was real based on this analysis and all of the Sanger data.

Table 3.16 Example read pair used for manual alignment with UCSC BLAT tool.

>forward GCTTACCTGCCGGCAGTTGTCTCCCACCTTGAAGATGGCTGCCTGCCAGGAGATCTTCTGG CCGTCGGCCTCCTGCGTGCTGCACTCATCCTCGGAGTCT
>reverse CGCTGTTAGCCTGTAGGCAGTGAGAAGCCCTCTGAGGGGGCCAGGGATGAGTCCTCTCAC ATGCCTGAAAGGAACCCCAGCAATCTTGGGGATCACTTCT

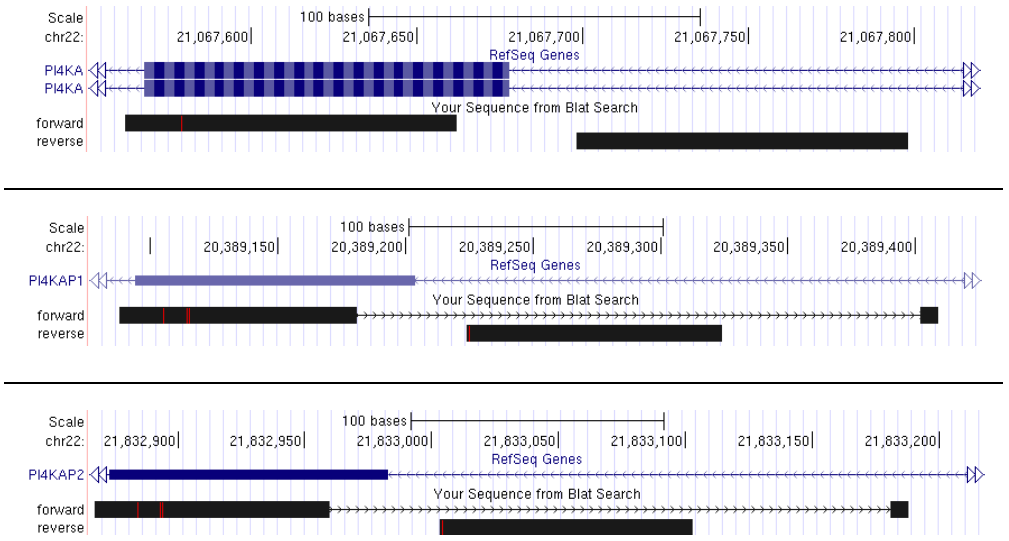


Figure 3.23: Alignment of read pairs from Table 3.16
BLAT alignment showing that read pairs mapped better with *PI4KA* compared with *PI4KAP1* and *PI4KAP2*. The read pairs are shown in UCSC Genome Browser. The black bars represent the aligned sequences and the red lines and gaps in the bars are regions that do not match the reference sequence.

3.2.4.2. Overview of *PI4KA*

PI4KA encodes a phosphatidylinositol kinase, in the AKT pathway. Variants in other phosphatidylinositol kinases from the pathway (*PIK3R2* and *PIK3CA*) have been implicated in PMG and other brain malformation phenotypes [184,185]. Variants in *AKT1* have been implicated in development disorders and overgrowth phenotypes [185,186]. There have been animal models developed to determine the effect of

deleterious variants in *PI4KA*. In the zebrafish model, down regulation of *PI4KA* was described as causing significant brain and fin abnormalities [187]. A conditional CRE/LOX [188] knockout mouse model was developed to study *PI4KA* as a potential Hepatitis C Virus drug target [183]. Although this kinase was not found to be a suitable drug target, the mouse model highlights the importance of this molecule in a functioning organism. When the homozygous knockout was induced all of the animals had to be euthanized due to their moribund state [183]. Since all the animals developed normally before the knockout was induced, it is difficult to surmise the effect of *PI4KA* on brain and development based on this model. Nevertheless, the two animal models are evidence of the potential importance of *PI4KA*.

3.2.4.3. Evidence of pathogenicity of *PI4KA* variants

PI4KA is in 22q11.2 locus, which is associated with PMG [189,190]. 22q11.2 is one of the most common clinically relevant microdeletions and has variable expressivity including DiGeorge, autism, schizophrenia and PMG. Whilst in most cases it is observed as a *de novo* variant, in certain cases it can be transmitted from unaffected carrier parents [7]. An initial hypothesis was that hemizigous variants in *PI4KA* might account for the PMG observed in certain cases of the deletion syndrome. BRC_232 is a patient in the Oxford Brain Abnormalities research cohort with PMG and a clinical diagnosis of deletion 22q11.2. The proband and mother were recruited into the project and exome sequenced to ascertain if there were any rare variants in *PI4KA* in the proband. WES was used as it was not feasible to accurately Sanger sequence all of *PI4KA* due to the two pseudo-genes. The 2.67Mb deletion from 18,894,835 to 21,561,514 encompassed the *PI4KA* locus and was visible in the exome sequencing but only examined closely because

of the *a priori* knowledge of its existence. There were no variants identified in *PI4KA* predicted to cause protein changes. A hemizygous variant would have been a simple answer to one of the phenotypes in a complex and variable deletion syndrome. The lack of variants in *PI4KA* in one patient does not rule out *PI4KA*'s involvement in 22q11 PMG cases. It is possible that there are other contributing factors that disrupt *PI4KA,* such as DNA methylation or haploinsufficiency which have been debated as causative factors in the various 22q11 syndromes [191,192]. Nevertheless, further functional evidence will be required to ascertain the pathogenicity of *PI4KA*.

To ascertain the function of the gene as it relates to human disease, a collaboration has been formed with Dr. Balla's research group at the National Institute of Child Health and Human Development, USA .The group was responsible for developing the zebrafish model described above. It was relatively clear to assess the stop gain variant as deleterious. The missense change was not as clear until it was revealed that the mutation is in the most critical part of the catalytic domain as part of the ATP binding pocket (from personal communication with Dr Balla). Currently, Dr Balla's group is developing the mutant protein and assessing its catalytic activity. Once developed and assessed, the protein will potentially be used for rescue experiments in a *PI4KA* knockout mouse.

This functional work will be vital in assessing the role of *PI4KA* in human brain abnormalities. However the identification of unrelated patients with similar phenotypes and biallelic mutations in *PI4KA* will be the only definitive way of proving the disease association.

3.2.5. Identification of *SOX11* variant in BRC_146

A significant result of this thesis was the identification of *SOX11* as a disease gene. Since there was no family history or consanguinity reported in BRC_146, a *de novo* variant was suspected and a trio approach with exome sequencing was undertaken. A *de novo* stop gain variant was identified in *SOX11* and which was validated by Sanger sequencing (Figure 3.24). Until recently this gene had not been associated with human disease [193]. The variant was at codon 29 of 441which suggests that it is likely to be a loss of function allele. The phenotype of BRC_115 was cerebellar vermis atrophy and microcephaly. The patient also presented with a particular head shape classified as trigonocephaly (triangle shaped forehead). *SOX11* encodes the SRY-BOX 11 transcription factor protein. It is part of a larger family of genes in which other members have been implicated in a range of human disease [194,195].

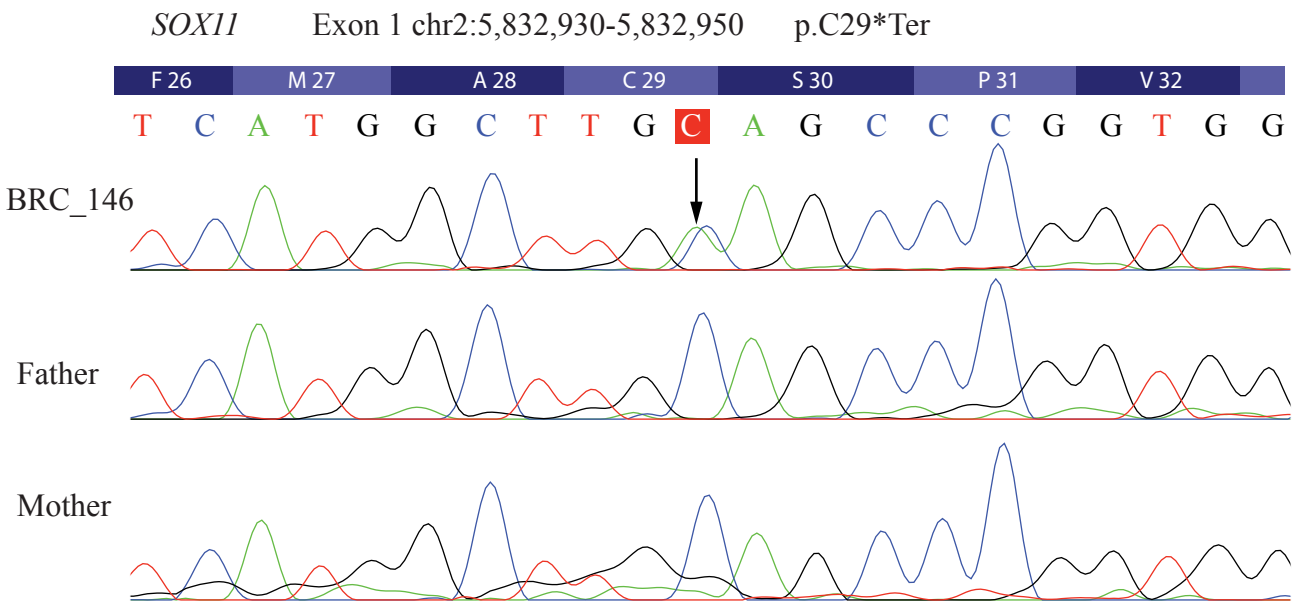


Figure 3.24: Sanger validation of *SOX11* variant in BRC_146
Sanger validation of *de novo* variant c.87C>A in *SOX11*. Arrow and red box indicates variant position.

3.2.5.1. Evidence of pathogenicity of *SOX11* variant

The phenotype to BRC_146 was similar to that observed in patients with variants in *SOX3,* which is associated with mental retardation and isolated growth hormone deficiency resulting in short stature and microcephaly [194]. The trigonocephlay phenotype of BRC_146 also matched a previous report of a single patient with a partial duplication of chromosome 17q and a terminal deletion of chromosome 2p encompassing SOX11 [196]. A mouse model demonstrated that *SOX11* is essential for embryonic and adult neurogenesis [197]. SiRNA interference of *SOX11* mRNA resulted in decreased expression of *TUBA1A* and *TUBB3* [198], genes associated with the neuronal migration disorder, Lissencephaly [199] and cortical malformations [200]. Based on this evidence, *SOX11* was being pursued as a potentially novel brain abnormality gene.

A group in Japan concurrently identified two *de novo* missense changes in *SOX11* with exome sequencing from a cohort from 92 Coffin-Siris syndrome (CSS) patients. The patients presented with a mild CSS phenotype with dimorphic facial features, microcephaly, growth deficiency, hypoplastic fifth toenails and mild intellectual disabilities [193]. BRC_146 did not present with typical CSS features, but did match some of the phenotype including hypoplastic toenails, in particular on the 3,4, and 5th toes. As a result of the Tsurusaki *et al* publication, it seems highly likely that this variant in *SOX11* is causative. Although *SOX11* is no longer novel, the gene and variant were found and prioritized as a strong candidate before the Japanese group published their findings. This highlights the success of the analytic approach and strategy of this project.

Chapter 4: Discussion

Chapter 4: Discussion

The aim of this thesis was to use next-generation sequencing to identify novel and known pathogenic mutations involved in brain abnormality phenotypes. A total of 23 patients were exome sequenced alongside 25 unaffected family members. The cohort was a combination of parent-parent-child trios, consanguineous families and families with multiple affected children.

This thesis has shown that the methods developed are effective for identifying variants of clinical relevance in highly heterogeneous brain abnormality syndromes. By identifying likely pathogenic variants in 7 of 18 families there was a definite positive clinical impact for patients. *PGAP3* is now a known brain abnormality disease gene and *PI4KA* is a potential disease gene. Methods developed in this project are expected to become a part of the clinical framework for using NGS in identifying the cause of brain abnormalities and other conditions.

4.1. Development of bioinformatics pipeline

A pipeline for analysis of WES data was developed as part of this thesis, which built on the pre-existing WTCHG bioinformatics pipeline. The utility of this pipeline and its analytical method was illustrated by discovery of pathogenic mutations in several patients with complex brain abnormality phenotypes, such as variants in novel disease gene *PGAP3* and the potential disease gene *PI4KA*.

4.1.1. Ingenuity Variant Analysis

A major aspect of the pipeline was the inclusion of Ingenuity Variant Analysis (IVA) as the primary VCF filtering tool. I undertook the assessment of IVA as part of my research. The expectation at the outset was that IVA was simply another unhelpful commercial product marketed for NGS data analysis that in reality would not provide any real benefit for research.

IVA proved, however, to be an effective tool in this project for candidate variant identification. It provides a simple, yet robust, method for filtering data from VCFs. Although there are other tools available that filter data, none of them combine access to a well-curated database alongside an intuitive user interface. Many of the traditional bioinformatics tools are UNIX command line based, and whilst these tools are effective, they are not readily accessible for many users. In IVA, it is relatively straightforward to filter for potential pathogenic variants in phenotypes of interest. Ease of access to the data facilitated rapid screening for variants in genes of interest from a diverse set of phenotypes with a number of inheritance models.

Another key feature of IVA is access to the Ingenuity knowledge base, which is a valuable database for analysing data in an unbiased manner. It is straightforward to change the Biological Context filtering to include a range of phenotypic terms. BRC_115 was an example of a family where an examination of a list of seizure related genes that include microcephaly as a phenotype, would not have identified *PCDH19*. Since it was easy to filter for different phenotypes, the variant was quickly identified. Evidence

associating the variant with the phenotype in IVA could always be easily accessed and scrutinized. The process of curating and examining evidence supporting a variant is a time consuming process. A recent study of clinical applicability of WGS found that it required a median of 54 minutes to assess the literature of each variant of interest [11]. By providing direct links to all of the sources supporting the association of variants, IVA decreases the overall time it takes to assess each variant.

A problem with using NGS in a clinical setting is how to properly integrate the expertise of the clinical scientists and referring clinicians with the more complicated data analysis aspects of NGS. An unrealistic goal is a single software package that could take the raw sequencing data and produce the likely pathogenic variants. By integrating IVA with the pipeline, it allows for NGS to be split in discrete steps or modules. Sequencing, mapping and variant calling can be done by the sequencing facility and the data analysis can be with IVA by experts in the particular phenotype being studied. Based on the work done as part of this thesis, IVA is now being used in other NGS projects focused on the diagnosis of genetic diseases at the WTCHG.

For all of the above reasons, IVA has been a powerful addition to WES pipeline at the WTCHG. Scripts that were developed as part of this pipeline to modify the VCFs for use in IVA have already been used in other on-going projects.

IVA is a commercial product and there is a cost associated with each sample. The pricing model works out to £88 per sample. NGS is an expensive method and adding an

additional per sample cost to a study would, therefore, decrease the total number of patient samples that could be sequenced. Although there would be less total samples for a given budget, all of the data would be scrutinized in an unbiased and efficient manner. More data is always useful, but if it is never properly analyzed, the sample may as well have not been sequenced. Since IVA provided a powerful means to analyze the data it has proved a worthwhile investment for this project.

4.1.2. Inclusion of the mitochondrial genome

In this project the mitochondrial genome was included in the target regions for exome sequencing. All of the mtDNA coverage was from off target sequencing reads and its inclusion could have introduced an additional source of false variants for the downstream analysis. Since the concordance comparison of WGS and WES mtDNA sequences reinforced the results from an earlier study[115], the inclusion of mtDNA sequences did not increase the total number of false variants. The mitochondrial genome is one of the most gene dense and clinical relevant parts of the human genome [114], and so should be included in the standard pipeline for WES.

4.1.3. Platypus minimum flank settings and variant annotation

A limitation of the current pipeline was the Platypus minimum flank setting that resulted in downstream variant annotation problems. The solution used for this thesis does not completely resolve these problems since every large multi nucleotide substitution is called separately. Variants in dbSNP are more likely to be annotated as single nucleotide

variants than multi-nucleotide variants. A comprehensive dbSNP reference list that included common variation typically found on the same allele would be one way to solve this issue. At this point, the best solution for the WES pipeline is to use Platypus with minimum flank set to zero. Identifying and resolving this issue has proved important for improving the efficiency of the pipeline.

4.1.4. Variant Databases

In this thesis, several variant databases such as OMIM, HGMD and ClinVar were used to see if variation had previously been reported in a clinical context. A limitation of these and other databases is that variants may have been misclassified as pathogenic in earlier studies. The estimates vary, but from 30 to 60% of variation classified as disease causing in mutation databases could be of uncertain significance [11,201]. A heterozygous variant was identified in this research in *PTCH1,* which a single report in HGMD suggested caused holoprosencephaly [202]. A further examination of the variant in ClinVar had it listed as non-pathogenic and a recent study also classified it as likely benign based on its frequency in healthy control populations [201]. When the literature was examined, it was clear even in the original paper that there was insufficient evidence supporting pathogenicity [202]. The variant is also in dbSNP and there have been no further cases of holoprosencephaly reported. This variant highlights challenges associated with variant databases. Until an international standard is established for pathogenic variant databases, the literature supporting variants will always need to be manually inspected.

As discussed above, IVA provided a means to easily assess the evidence supporting the pathogenicity of a variant. Similar to most variant identification software, IVA is dependent on the accuracy of the databases and transcript prediction sources being used to annotate the variants. If there are errors in the datasets, IVA has no method to identify miss-annotated variation. For this reason, the annotation of candidate variants must be inspected to ensure that there are no errors.

4.2. Clinical application of NGS

This project was successful in identifying likely pathogenic variants in 7 of 18 families. I identified five variants in genes that had previously been implicated in brain malformation syndromes, a similar result to other NGS studies of heterogeneous conditions [4,14,51,55]. Since these conditions can be caused by multiple genes with phenotyptic variability, NGS studies will frequently identify variants in known disease genes [13]. Information on these variants has been provided to the Oxford Regional Genetics Laboratory for subsequent clinical validation and eventual diagnosis for the patients.

4.2.1. Comparative sensitivity of Sanger sequencing and NGS

An inclusion criterion for this project was that patients had been screened for variants in *ASPM* but variants were missed by Sanger sequencing in BRC_90. The WGS500 (manuscript in preparation) and other NGS studies [60,93] have shown there is an appreciable false negative rate when screening genes by traditional Sanger sequencing.

This study reaffirms the error rate seen with Sanger sequencing and provides another compelling argument for migration to NGS technologies.

There are several reasons why Sanger sequencing can miss a variant, a SNP on one allele in the primers sequences can lead to mono-allelic amplification thus masking a variant [203,204], or the heterozygous peak could be missed because they are difficult to call. Clinical Sanger sequencing testing can be automated which can also result in variants being missed if the software fails to recognize the smaller heterozygous peak. Although variant calling with NGS can also be automated, heterozygosity can be detected even with low minor allele frequencies. An advantage of NGS is that because each read represents a single molecule, this can result in a high sensitivity when looking for variants with low allelic ratios that are characteristic of somatic mosaicism. It is likely that such variants are often missed by Sanger sequencing [93,205]. With the decreasing cost of sequencing [34], WES and WGS will likely become the standard clinical method for testing for variants in known disease genes. As this thesis demonstrated, NGS is an effective method for diagnosing genetic disease conditions.

4.2.2. Previous clinical testing of known genes

A second inclusion criterion of this project was that the patient's phenotype was not characteristic of a known syndromic brain abnormality disorder. It emerged however that brain abnormality phenotypes are too heterogeneous and variable for clinicians to have in depth knowledge about all conditions. An expert in the genetics of X-linked learning disability syndromes with experience of interpreting *CASK* variants and their associated

phenotypes, might have ordered a single gene test for BRC_59 and BRC_16 instead of WES. However a different clinician might have focused on the cerebellar hypoplasia reported in BRC_16 and concluded that the patient phenotype was more characteristic of another gene with variable phenotypes that include hypoplasia, such *TSEN54* [206], *OPHN1*[207] or *POMT1*[208].

An issue with clinical tests for genetically heterogeneous conditions is that there are several likely genes that could explain the phenotype and a genetic diagnosis will always be delayed by the cost of the tests. Currently, each test is undertaken sequentially in order to save funds, which can lead to a long wait for the patient and their family. This can lead to a diagnostic odyssey involving test after test that may or may not find the causative gene. A genetic diagnosis must be made in a timely manner for it to have the greatest impact. For example, parents might decide not to have another child if they do not know the recurrence risk of the genetic disorder and if a diagnosis takes several years, the parents might then be too old to have another child. Currently, it only takes 2 to 4 single gene tests for the cumulative cost to exceed that of WES [209]. Testing both of the two most common microcephaly genes *ASPM* and *WDR62* in the OUH Regional Genetics Laboratory, would cost £1214 (based on 2014 internal NHS costs) indicating that it would have been cost effective to provide WES as a first line clinical test. In the case of heterogeneous conditions, such as brain abnormalities, NGS is a solution that can provide a timely diagnosis and potentially save funds.

4.2.3. Clinically relevant findings and the benefit of genetic diagnosis

In the 7 of the 18 families, I identified a genetic diagnosis, representing a significant clinical benefit for the families involved. Genetic counselling will advise the parents in regards to the recurrence risk of having another child. Provided the four *de novo* variants are causative, and there is no germline mosaicism in the parents [41], the recurrence risk will be very low. In the case of the autosomal recessive conditions, the parents can explore the option of prenatal testing or pre-implantation genetic diagnosis (PGD). However, for many of the patients, there will be no immediate clinical benefit because there are no treatments currently available for their major phenotypes.

A diagnosis of a GPI disorder in the BRC_41 and siblings has potentially improved treatment. Early evidence has shown that in some patients with GPI disorders, pyridoxine (vitamin B6) is an effective treatment for seizures [73,210]. The family is also seeking to have PGD to potentially have a healthy child without *PGAP3* variants. The diagnosis was timely because if the mother had been 1 year older, she would not have qualified for NHS funded PGD [211].

By identifying further disease causing variants in these genes, it might improve the overall understanding of the phenotypes. *CDK5RAP2* was first identified as a disease-causing gene in 2005, but due to the rarity of the condition the phenotype outside of the primary microcephaly is still not fully defined. Aspects of BRC_81's non-CNS phenotype are being compared with a knockout *CDK5RAP2* mouse model being

developed in Germany. This collaboration will improve the overall understanding of the phenotype, which could in turn benefit future patients.

Discovery of novel brain abnormality disease genes with NGS will provide insight into brain development and disease mechanisms [7]. A greater understanding of the mechanisms that underlie many brain abnormalities will, over the long term, improve patient care and treatment.

4.2.4. Reverse phenotyping

PCDH19 was a surprising result in BRC_115 as it had only been implicated in epilepsy in females with mental retardation (EFMR). Although the patient had significant developmental delay and seizures, microcephaly was originally thought to be the primary diagnosis. The genetic diagnosis was a case of reverse phenotyping [13,212], where the identified genetic variant has helped clinicians in making a diagnosis. *CASK* variants identified in BRC_59 and BRC_16 are also examples where NGS diagnosed the patient. These are further examples of how the unbiased nature of NGS is clinically beneficial for patients and their caregivers [13].

4.3. Novel genes for brain abnormality phenotypes

In this project, I identified variants in the novel disease genes *PGAP3* and *PI4KA*. Through the work done as part of this thesis, *PGAP3* is now a known disease gene listed in OMIM as associated with hyperphosphatasia with mental retardation syndrome. *PI4KA*

has been identified as a potential disease gene and functional testing will determine if it is truly responsible for the phenotype shown in the patients.

4.3.1. PGAP3

The experimental work undertaken in identifying *PGAP3* and the subsequent collaborative efforts highlight the efficacy of the research approach of this thesis in identifying a novel brain abnormality gene. The variants identified, alongside those previously reported in *PGAP2* [174,178] and PGAP1 [213], confirm the importance of the fatty acid remodelling steps of the GPI pathway for normal neurological development.

The combination of autozygosity mapping and WES, highlight the advantage of NGS for disease gene discovery in consanguineous families with heterogeneous brain abnormalities. The single 13.8 Mb region of autozygosity identified in the three affected siblings contains 374 annotated genes. In hindsight, *PGAP3* is a good candidate gene, but there were multiple genes in pathways which seemed as likely to cause the phenotype, such as *MAPT*, the microtubule associated protein. Since a significant number of the genes are expressed in the developing human brain [8,9], it was not easy to narrow the list of candidates. Linkage analysis with multiple affected families could have narrowed the candidate region, but there was marked phenotypic variability in the patients with *PGAP3* variants, and it would have been difficult to build a large cohort. DNA from the unaffected double first cousins could have potentially narrowed the region, but was not available since the extended family still resides in Pakistan. The high serum alkaline phosphatase is a clear indicator of GPI deficiency and could also have pointed to *PGAP3*

as being the most likely gene within the autozygous region. However, this biochemical result was only noted retrospectively once the variant had been found. For all of the above reasons, it would have been a tremendous undertaking to identify *PGAP3* without the use of NGS.

One differing feature among the patients identified with *PGAP3* variants, is the postnatal development of head circumference. In the Oxford family, all affected individuals developed microcephaly, whereas the OFC of the other individuals is in the mean age adjusted range. This is not inconsistent with GPI disorders [214], as there is marked variability regarding postnatal growth and OFC in the groups of affected individuals with *PIGV*, *PIGO*, and *PGAP2* variants. A multi genic effect was considered in the Oxford family but there were no other candidate genes that would explain the microcephaly. *PGAP3* is another example of the phenotypic variability seen in brain abnormality syndromes. Despite all the patients having variants in *PGAP3*, the location of the variants, or an epigenetic effect, was causing a variable phenotype. This is further evidence of the advantage of the unbiased NGS approach in diagnosing complex heterogeneous conditions.

4.3.2. PI4KA

PI4KA was an example of a potential novel disease gene identified with this research approach. Significant functional work must be done before the variants can be properly classified as pathogenic. Even with animal model work, finding unrelated patients with similar variants and phenotypes would be the strongest indicator of *PI4KA*'s role in

human disease. Attempts have already been made to find further patients by contacting other research groups focused on severe congenital abnormalities. Through collaboration with Dr Balla further patients may be found, as occurred in the case of *PGAP3*, but this is unlikely because based on the animal model work conducted by his lab group, Dr Balla speculates that most *PI4KA* variants would be embryonically lethal.

Proving the pathogenicity of these variants will have an immediate impact on the family. Three terminated pregnancies due to severe congenital abnormalities is a significant psychological burden on the family. Attempting to have a healthy child through PGD is likely the best option for this family. For PGD to be covered by the NHS there is a requirement of significant proof of pathogenicity, and shared variants among three affected fetuses that were exome sequenced will be insufficient. Functional work might be enough, but it could take a significant amount of time for Dr Balla's group to complete the lab work. There are no further siblings or cousins to test for segregation.

One of the variants identified in *PI4KA* was in a segmental duplication region. Two pseudo-gene copies of *PI4KA* prevented efficient Sanger sequencing of the missense change in exon 48 and would likely interfere with the sequencing of several of the other exons. A traditional disease gene approach using SNP arrays, could have implicated the region as harbouring the disease gene based on the identical by descent regions in the three fetuses. Subsequent Sanger sequencing of *PI4KA* as a candidate gene would have been severely hampered by the noisy chromatogram traces due to the segmental duplications. In WES, the region was covered by an average 82.4 reads in the family.

Since the Stampy [97] mapping algorithm was able to effectively differentiate between the segmental duplication reads, the missense variant was identified. NGS was a superior method for sequencing *PI4KA*. It is unlikely that the missense variant would have been found with traditional disease gene identification methods.

Since there was limited DNA in this family, WGS might have been the best approach. Instead of screening only one of the fetuses for CNVs with SNP arrays followed by WES, WGS of all of the siblings could have been used to screen for CNVs and saved the aliquot of DNA used in the arrays [215]. Although CNVs can also be detected from exome data [216], it has the same issue as the sequencing, only the exonic regions are screened. It is also possible in this family that genome sequencing would identify stronger candidates.

4.4. Future

To date, the majority of Mendelian disease causing variants have been reported in exonic regions [5,12]. This will likely change over the next five to ten years as WGS becomes the standard for next-generation sequencing studies.

Since many genes are implicated in brain abnormalities, an NGS approach with panel testing could provide a solution. Limitations of this approach include the time that it takes to develop and validate the panel. Many brain abnormalities are extremely rare and the high rate of publications in this field means it would be difficult to include all known

genes on a single panel. For these reasons, WES is currently the best option for diagnosing a heterogeneous condition such as brain abnormalities.

Since the cost to sequence a genome is decreasing [34], WGS will soon become a realistic option for rare disease gene research. Studying brain abnormalities with WGS will have the advantages of the NGS approach but will also capture other regions of the genome that could harbour potentially pathogenic variants. Even with the newest exome capture kits, there are still some exonic regions known to cause human disease that are not well covered [5]. This was shown by a recent study of patients with severe intellectual disability who previously been screened by WES. WGS found additional conclusive causative variants in 21 of 50 patients [16]. It is clear that WGS will provide a genetic diagnosis in many cases that are missed by WES. In this project it is possible that in some of the cases without candidates, a WGS approach might identify pathogenic variants.

As illustrated in this thesis, new sequencing technologies and analytical approaches for genome annotation are transforming our ability to understand rare disease mechanisms. In addition, they provide unique platforms for improving patient care for those with rare diseases. The speed and efficiency with which we were able to dissect the genetic basis of a wide set of brain abnormalities indicates how powerful these tools have become. The future is likely to see the generation of more comprehensive databases of causative variants for these diseases, improved diagnostics and, ultimately, novel therapies.

Reference List

1. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, et al. (2013) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. science 291: 1304-1351.
4. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14: 681-691.
5. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. European Journal of Human Genetics 20: 490-497.
6. Howard MF, Murakami Y, Pagnamenta AT, Daumer-Haas C, Fischer B, et al. (2014) Mutations in PGAP3 Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation. The American Journal of Human Genetics 94: 278-287.
7. Dyment DA, Sawyer SL, Chardon JW, Boycott KM (2013) Recent advances in the genetic etiology of brain malformations. Current neurology and neuroscience reports 13: 1-10.
8. Naumova OY, Lee M, Rychkov SY, Vlasova NV, Grigorenko EL (2013) Gene expression in the human brain: the current state of the study of specificity and spatiotemporal dynamics. Child Dev 84: 76-88.
9. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. Nature 478: 483-489.
10. Consortium GP (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
11. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, et al. (2014) Clinical interpretation and implications of whole-genome sequencing. JAMA 311: 1035-1045.
12. Samuels DC, Han L, Li J, Quanghu S, Clark TA, et al. (2013) Finding the lost treasures in exome sequencing data. Trends Genet 29: 593-599.
13. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. genome 11: 64.
14. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.
15. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. (2010) Making a definitive diagnosis: successful clinical application of whole exome

sequencing in a child with intractable inflammatory bowel disease. Genetics in Medicine 13: 255-262.
16. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, et al. (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature.
17. Jacob HJ, Abrams K, Bick DP, Brodie K, Dimmock DP, et al. (2013) Genomics in clinical practice: lessons from the front lines. Science translational medicine 5: 194cm195-194cm195.
18. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74: 5463-5467.
19. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, et al. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.
20. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, et al. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1: 263-273.
21. Collins FS (1992) Positional cloning: let's not call it reverse anymore. Nature genetics 1: 3-6.
22. Wolff RK (1997) Positional cloning: a review and perspective. Drug development research 41: 129-141.
23. Ballabio A (1993) The rise and fall of positional cloning? Nature genetics 3: 277-279.
24. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, et al. Cloning the gene for the inherited disorder chronic granulomatous disease on the basis of its chromosomal location; 1986. Cold Spring Harbor Laboratory Press. pp. 177-183.
25. Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, et al. (1986) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene.
26. Weinberg RA, Rapaport JM, Albert DM (1986) A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 323: 643.
27. Kerem B-s, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073-1080.
28. Riordan JR, Rommens JM, Kerem B-s, Alon N, Rozmahel R, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245: 1066-1073.
29. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236: 1567-1570.
30. Jackson AP, McHale DP, Campbell DA, Jafri H, Rashid Y, et al. (1998) Primary autosomal recessive microcephaly (MCPH1) maps to chromosome 8p22-pter. Am J Hum Genet 63: 541-546.
31. Jackson AP, Eastwood H, Bell SM, Adu J, Toomes C, et al. (2002) Identification of microcephalin, a protein implicated in determining the size of the human brain. Am J Hum Genet 71: 136-142.

Reference List

32. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics 33: 228-237.
33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.
34. Wetterstrand K (2014) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcosts.
35. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.
36. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.
37. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42: 30-35.
38. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, et al. (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. New England Journal of Medicine 362: 1181-1191.
39. Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, et al. (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS genetics 6: e1000991.
40. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nature Reviews Genetics 8: 610-618.
41. Veltman JA, Brunner HG (2012) De novo mutations in human genetic disease. Nat Rev Genet 13: 565-575.
42. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nature 201.
43. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
44. Neale BM, Kou Y, Liu L, Ma'Ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242-245.
45. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, et al. (2010) A de novo paradigm for mental retardation. Nature genetics 42: 1109-1112.
46. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences 107: 961-968.
47. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, et al. (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nature genetics 42: 483-485.
48. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics 42: 790-793.
49. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics 43: 585-589.

Reference List

50. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, et al. (2011) Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nature genetics 43: 864-868.
51. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, et al. (2012) Diagnostic exome sequencing in persons with severe intellectual disability. New England Journal of Medicine 367: 1921-1929.
52. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, et al. (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. The Lancet 380: 1674-1682.
53. Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, et al. (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 42: 827-829.
54. Becker J, Semler O, Gilissen C, Li Y, Bolz HJ, et al. (2011) Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. The American Journal of Human Genetics 88: 362-371.
55. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, et al. (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478: 57-63.
56. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12: 745-755.
57. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, et al. (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Science translational medicine 3: 65ra64-65ra64.
58. Feero WG, Guttmacher AE, Dietz HC (2010) New therapeutic approaches to mendelian disorders. New England Journal of Medicine 363: 852-863.
59. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences 106: 19096-19101.
60. Bonnefond A, Durand E, Sand O, De Graeve F, Gallina S, et al. (2010) Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. PloS one 5: e13630.
61. Illumina (2014) HiSeq X Ten for Population-Scale Sequencing - World's First $1000 Human Genome. http://www.illumina.com/systems/hiseq-x-sequencing-system/system.ilmn.
62. Moreau Y, Tranchevent L-C (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics 13: 523-536.
63. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, et al. (2006) Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. American journal of medical genetics Part A 140: 2063-2074.
64. Shoukier M, Klein N, Auber B, Wickert J, Schröder J, et al. (2013) Array CGH in patients with developmental delay or intellectual disability: are there phenotypic clues to pathogenic copy number variants? Clinical genetics 83: 53-65.
65. Stegmann A, Jonker L, Engelen J (2008) Prospective screening of patients with unexplained mental retardation using subtelomeric MLPA strongly increases the

detection rate of cryptic unbalanced chromosomal rearrangements. European journal of medical genetics 51: 93-105.
66. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, et al. (2009) Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. Genetics in Medicine 11: 139-146.
67. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, et al. (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. The American Journal of Human Genetics 86: 749-764.
68. OMIM OMIM (2014) Clinical Synopsis Advanced Search. http://omim.org/search/advanced/clinicalSynopsis.
69. Kaindl AM (2014) Autosomal recessive primary microcephalies (MCPH). European journal of paediatric neurology: EJPN: official journal of the European Paediatric Neurology Society.
70. Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, et al. (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467: 207-210.
71. Rees MI (2010) The genetics of epilepsy—the past, the present and future. Seizure 19: 680-683.
72. Nicholas AK, Khurshid M, Désir J, Carvalho OP, Cox JJ, et al. (2010) WDR62 is associated with the spindle pole and is mutated in human microcephaly. Nature genetics 42: 1010-1014.
73. Kuki I, Takahashi Y, Okazaki S, Kawawaki H, Ehara E, et al. (2013) Vitamin B6–responsive epilepsy due to inherited GPI deficiency. Neurology 81: 1467-1469.
74. Woods CG, Bond J, Enard W (2005) Autosomal recessive primary microcephaly (MCPH): a review of clinical, molecular, and evolutionary findings. Am J Hum Genet 76: 717-728.
75. O'Brien DP, Campbell KA, Morken NW, Bair RJ, Heath EM (2001) Automated Nucleic Acid Purification for Large Samples. Journal of the Association for Laboratory Automation 6: 67-70.
76. Thermo_Scientific (2014) NanoDrop Technical support bulletin T009 (2007). http://www.bio.davidson.edu/projects/gcat/protocols/NanoDrop_tip.pdf.
77. O'neill M, McPartlin J, Arthure K, Riedel S, McMillan ND (2011) Comparison of the TLDA with the Nanodrop and the reference Qubit system. Sensors & Their Applications Xvi 307.
78. Life_Technologies (2010) Accurate and sensitive quantitation of nucleic acids, even at low concentrations - Comparison of the Qubit Quantitation Platform with Spectrophotometry. http://www.lifetechnologies.com/content/dam/LifeTech/migration/en/filelibrary/support/bioprobes/bioprobes-62.par.14365.file.dat/bioprobes-62-qubit.pdf.
79. Illumina (2009) Infinium HD Assay Ultra Protocol Guide. http://supportres.illumina.com/documents/myillumina/67f59f89-51ee-44d6-b1bb-a53dcb5bd01e/infinium_hd_ultra_user_guide_11328087_revb.pdf.

Reference List

80. Ritchie ME, Liu R, Carvalho BS, Australia, New Zealand Multiple Sclerosis Genetics C, et al. (2011) Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics 12: 68.
81. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, et al. (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38: e105.
82. Pagnamenta AT, Murray JE, Yoon G, Akha ES, Harrison V, et al. (2012) A Novel Nonsense CDK5RAP2 Mutation in a Somali Child With Primary Microcephaly and Sensorineural Hearing Loss. American Journal of Medical Genetics Part A 158A: 2577-2582.
83. Pagnamenta AT, Wing K, Sadighi Akha E, Knight SJ, Bolte S, et al. (2009) A 15q13.3 microdeletion segregating with autism. Eur J Hum Genet 17: 687-692.
84. Knight SJ, Yau C, Clifford R, Timbs AT, Sadighi Akha E, et al. (2012) Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with B-cell chronic lymphocytic leukemia. Leukemia 26: 1564-1575.
85. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42: D986-992.
86. Seelow D, Schuelke M, Hildebrandt F, Nurnberg P (2009) HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res 37: W593-599.
87. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
88. Seelow D, Schuelke M, Hildebrandt F, Nurnberg P (2009) HomozygosityMapper Technical documentation. http://www.homozygositymapper.org/technical_documentation.html - data_analysis.
89. Seelow D, Schwarz JM, Schuelke M (2008) GeneDistiller--distilling candidate genes from linkage intervals. PLoS One 3: e3874.
90. McKusick-Nathans Institute of Genetic Medicine JHU Online Mendelian Inheritance in Man, OMIM®.
91. Consortium GR (2014) Human Genome Assembly Data GRCh37.p13. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/.
92. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, et al. (2008) Runs of homozygosity in European populations. American Journal of Human Genetics 83: 359-372.
93. Pagnamenta AT, Lise S, Harrison V, Stewart H, Jayawant S, et al. (2012) Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. Journal of Human Genetics 57: 70-72.
94. Martin HC, Kim GE, Pagnamenta AT, Murakami Y, Carvill GL, et al. (2014) Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. Hum Mol Genet.
95. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome research 12: 656-664.

Reference List

96. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
97. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936-939.
98. Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm.
99. Picard (2014) Picard command-line tools. 1.92 ed. http://picard.sourceforge.net.
100. Li H (2014) Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. arXiv preprint arXiv:14040929.
101. Wang S, Xing J (2013) A Primer for Disease Gene Prioritization Using Next-Generation Sequencing Data. Genomics & informatics 11: 191-199.
102. Illumina (2014) Infinium® Genotyping Data Analysis - A guide for analyzing Infinium genotyping data using the GenomeStudio® Genotyping Module. http://res.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf.
103. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
104. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
105. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91: 839-848.
106. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, et al. (2012) Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28: 2678-2679.
107. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet.
108. Rimmer A, Mathieson I, Lunter G, McVean G (2012) Platypus: An Integrated Variant Caller. http://www.well.ox.ac.uk/platypus.
109. Cossins J, Belaya K, Hicks D, Salih MA, Finlayson S, et al. (2013) Congenital myasthenic syndromes due to mutations in ALG2 and ALG14. Brain: awt010.
110. Fernandez-Mercado M, Burns A, Pellagatti A, Giagounidis A, Germing U, et al. (2013) Targeted resequencing analysis of 25 genes commonly mutated in myeloid disorders in del (5q) myelodysplastic syndromes. haematologica: haematol. 2013.086686.
111. Tatton-Brown K, Seal S, Ruark E, Harmer J, Ramsay E, et al. (2014) Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nature genetics.
112. Broad_Institute (2014) GATK - How should I interpret VCF files produced by the GATK? http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk.
113. Broad_Institute (2014) Introduction to the GATK Genome Sequencing Analysis BestPractices. http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-0B-Intro_to_GATK.pdf.
114. (2013) MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org,.

Reference List

115. Picardi E, Pesole G (2012) Mitochondrial genomes gleaned from human whole-exome sequencing. Nature Methods 9: 523-524.
116. Ingenuity (2014) Predicted Deleterious Filter. http://ingenuity.force.com/variants/VariantTutorials?id=kA250000000TNUFCA4.
117. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. Nature 437: 1032-1037.
118. Ingenuity (2014) Ingenuity Knowledge Base. http://www.ingenuity.com/science/knowledge-base.
119. Escayg A, MacDonald BT, Meisler MH, Baulac S, Huberfeld G, et al. (2000) Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+ 2. Nature genetics 24: 343-345.
120. Johnston JJ, Gropman AL, Sapp JC, Teer JK, Martin JM, et al. (2012) The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. Am J Hum Genet 90: 295-300.
121. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, et al. (2013) ENCODE Data in the UCSC Genome Browser: year 5 update. Nucleic Acids Research 41: D56-D63.
122. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. (2013) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research: gkt1113.
123. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, et al. (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genetics in Medicine 15: 565-574.
124. ACMG (2014) ACMG Updates Recommendation on "Opt Out" for Genome Sequencing Return of Results
. https://http://www.acmg.net/docs/Release_ACMGUpdatesRecommendations_final.pdf.
125. Wolf SM, Annas GJ, Elias S (2013) Respecting Patient Autonomy in Clinical Genomics: New Recommendations on Incidental Findings Go Astray. Science (New York, NY) 340: 1049.
126. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.
127. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 12: R112.
128. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Research 12: 996-1006.
129. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.
130. NCBI (2014) NCBI database of genetic variation dbSNP (138). http://www.ncbi.nlm.nih.gov/projects/SNP/.
131. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003-1007.

Reference List

132. Smit A, Hubley, R & Green, P. (1996-2010) RepeatMasker Open-3.0. http://www.repeatmasker.org.
133. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
134. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.
135. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, et al. (2012) Primer3-new capabilities and interfaces. Nucleic Acids Research 40.
136. Jurka J (2000) Repbase Update - a database and an electronic journal of repetitive elements. Trends in Genetics 16: 418-420.
137. Birch DE, Kolmodin L, Wong J, Zangenberg G, Zoccoli M (1996) Simplified hot start PCR. Nature 381: 445-446.
138. Hanke M, Wink M (1994) Direct DNA sequencing of PCR-amplified vector inserts following enzymatic degradation of primer and dNTPs. Biotechniques 17: 858-860.
139. Applied_Biosystems (2002) Applied Biosystems 3730/3730xl DNA Analyzers Sequencing Chemistry Guide. http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_041258.pdf.
140. Applied_Biosystems (2009) DNA Sequencing by Capillary Electrophoresis Applied Biosystems Chemistry Guide Second Edition. http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_041003.pdf.
141. McCarthy C (1996) Chromas: version 2.0. Technelysium PTY, Australia.
142. Griekspoor A, Groothuis T (2005) 4Peaks: a program that helps molecular biologists to visualize and edit their DNA sequence files v1. 7.
143. Gordon D, Green P (2013) Consed: a graphical editor for next-generation sequencing. Bioinformatics 29: 2936-2937.
144. Stephens M, Sloan JS, Robertson P, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nature genetics 38: 375-381.
145. Baker M (2010) Next-generation sequencing: adjusting to data overload. Nature Methods 7: 495-499.
146. Miyake A, Itoh N (2013) Fgf22 regulated by Fgf3/Fgf8 signaling is required for zebrafish midbrain development. Biology open 2: 515-524.
147. Butterworth RF (2011) Neuroinflammation in acute liver failure: mechanisms and novel therapeutic targets. Neurochemistry international 59: 830-836.
148. Uehara M, Yashiro K, Mamiya S, Nishino J, Chambon P, et al. (2007) CYP26A1 and CYP26C1 cooperatively regulate anterior–posterior patterning of the developing brain and the production of migratory cranial neural crest cells in the mouse. Developmental biology 302: 399-411.
149. Roessler E, Ouspenskaia MV, Karkera JD, Vélez JI, Kantipong A, et al. (2008) Reduced NODAL signaling strength via mutation of several pathway members including FOXH1 is linked to human heart defects and holoprosencephaly. The American Journal of Human Genetics 83: 18-29.

Reference List

150. Kim Y, Sharov AA, McDole K, Cheng M, Hao H, et al. (2011) Mouse B-type lamins are required for proper organogenesis but not by embryonic stem cells. Science 334: 1706-1710.
151. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.
152. Postel J, Reynolds J (1985) File transfer protocol.
153. Wolf MT, Mucha BE, Attanasio M, Zalewski I, Karle SM, et al. (2003) Mutations of the Uromodulin gene in MCKD type 2 patients cluster in exon 4, which encodes three EGF-like domains. Kidney international 64: 1580-1587.
154. Smith GD, Robinson C, Stewart AP, Edwards EL, Karet HI, et al. (2011) Characterization of a recurrent in-frame UMOD indel mutation causing late-onset autosomal dominant end-stage renal failure. Clinical Journal of the American Society of Nephrology 6: 2766-2774.
155. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069-2070.
156. Illumina (2014) Illumina VariantStudio Data Analysis Software. http://res.illumina.com/documents/products/datasheets/datasheet_illumina_variantstudio_software.pdf.
157. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, et al. (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10: 294-300.
158. Najm J, Horn D, Wimplinger I, Golden JA, Chizhikov VV, et al. (2008) Mutations of CASK cause an X-linked brain malformation phenotype with microcephaly and hypoplasia of the brainstem and cerebellum. Nature genetics 40: 1065-1067.
159. Moog U, Kutsche K, Kortum F, Chilian B, Bierhals T, et al. (2011) Phenotypic spectrum associated with CASK loss-of-function mutations. J Med Genet 48: 741-751.
160. Truong HT, Dudding T, Blanchard CL, Elsea SH (2010) Frameshift mutation hotspot identified in Smith-Magenis syndrome: case report and review of literature. Bmc Medical Genetics 11.
161. ClinVar (2014) CASK:c.1639C>T (p.Gln547Ter) AND Mental retardation and microcephaly with pontine and cerebellar hypoplasia. http://www.ncbi.nlm.nih.gov/clinvar/RCV000022833/.
162. Bond J, Roberts E, Springell K, Lizarraga S, Scott S, et al. (2005) A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. Nature genetics 37: 353-355.
163. Hassan MJ, Khurshid M, Azeem Z, John P, Ali G, et al. (2007) Previously described sequence variant in CDK5RAP2 gene in a Pakistani family with autosomal recessive primary microcephaly. BMC medical genetics 8: 58.
164. Depienne C, Bouteiller D, Keren B, Cheuret E, Poirier K, et al. (2009) Sporadic Infantile Epileptic Encephalopathy Caused by Mutations in PCDH19 Resembles Dravet Syndrome but Mainly Affects Females. Plos Genetics 5.
165. Depienne C, Trouillard O, Bouteiller D, Gourfinkel-An I, Poirier K, et al. (2011) Mutations and Deletions in PCDH19 Account for Various Familial or Isolated Epilepsies in Females. Human Mutation 32: E1959-E1975.

Reference List

166. Kwong AK, Fung CW, Chan SY, Wong VC (2012) Identification of SCN1A and PCDH19 mutations in Chinese children with Dravet syndrome. PLoS One 7: e41802.
167. Terracciano A, Specchio N, Darra F, Sferra A, Bernardina BD, et al. (2012) Somatic mosaicism of PCDH19 mutation in a family with low-penetrance EFMR. Neurogenetics 13: 341-345.
168. Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer‐Sawatzki H, et al. (2011) On the sequence‐directed nature of human gene mutation: The role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Human mutation 32: 1075-1099.
169. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods 7: 575-576.
170. Fujita M, Kinoshita T (2012) GPI-anchor remodeling: Potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids 1821: 1050-1058.
171. McKean DM, Niswander L (2012) Defects in GPI biosynthesis perturb Cripto signaling during forebrain development in two new mouse models of holoprosencephaly. Biol Open 1: 874-883.
172. Park S, Lee C, Sabharwal P, Zhang M, Meyers CL, et al. (2013) GDE2 promotes neurogenesis by glycosylphosphatidylinositol-anchor cleavage of RECK. Science 339: 324-328.
173. Wei X, Xu R, Yang Z, Li Z, Liao Y, et al. (2012) Novel uromodulin mutation in familial juvenile hyperuricemic nephropathy. American journal of nephrology 36: 114-120.
174. Hansen L, Tawamie H, Murakami Y, Mang Y, Ur Rehman S, et al. (2013) Hypomorphic Mutations in PGAP2, Encoding a GPI-Anchor-Remodeling Protein, Cause Autosomal-Recessive Intellectual Disability. Am J Hum Genet 92: 575-583.
175. Kvarnung M, Nilsson D, Lindstrand A, Korenke GC, Chiang SC, et al. (2013) A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. Journal of Medical Genetics.
176. Nozaki M, Ohishi K, Yamada N, Kinoshita T, Nagy A, et al. (1999) Developmental abnormalities of glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/loxP system. Lab Invest 79: 293-299.
177. Murakami Y, Kanzawa N, Saito K, Krawitz PM, Mundlos S, et al. (2012) Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome. J Biol Chem 287: 6318-6325.
178. Krawitz PM, Murakami Y, Riess A, Hietala M, Kruger U, et al. (2013) PGAP2 Mutations, Affecting the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation Syndrome. Am J Hum Genet 92: 584-589.
179. Tashima Y, Taguchi R, Murata C, Ashida H, Kinoshita T, et al. (2006) PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Mol Biol Cell 17: 1410-1420.

Reference List

180. Maeda Y, Tashima Y, Houjou T, Fujita M, Yoko-o T, et al. (2007) Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol Biol Cell 18: 1497-1506.
181. Shipman KE, Holt AD, Gama R (2013) Interpreting an isolated raised serum alkaline phosphatase level in an asymptomatic patient. Bmj-British Medical Journal 346.
182. Möller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17: 646-653.
183. Vaillancourt FH, Brault M, Pilote L, Uyttersprot N, Gaillard ET, et al. (2012) Evaluation of phosphatidylinositol-4-kinase IIIα as a hepatitis C virus drug target. Journal of virology 86: 11595-11607.
184. Rivière J-B, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, et al. (2012) De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature genetics 44: 934-940.
185. Orloff MS, He X, Peterson C, Chen F, Chen J-L, et al. (2013) Germline PIK3CA and AKT1 Mutations in Cowden and Cowden-like Syndromes. The American Journal of Human Genetics 92: 76-80.
186. Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, et al. (2011) A mosaic activating mutation in AKT1 associated with the Proteus syndrome. New England Journal of Medicine 365: 611-619.
187. Ma H, Blake T, Chitnis A, Liu P, Balla T (2009) Crucial role of phosphatidylinositol 4-kinase IIIα in development of zebrafish pectoral fin is linked to phosphoinositide 3-kinase and FGF signaling. Journal of cell science 122: 4303-4310.
188. Orban PC, Chui D, Marth JD (1992) Tissue-and site-specific DNA recombination in transgenic mice. Proceedings of the National Academy of Sciences 89: 6861-6865.
189. Robin NH, Taylor CJ, McDonald‐McGinn DM, Zackai EH, Bingham P, et al. (2006) Polymicrogyria and deletion 22q11. 2 syndrome: window to the etiology of a common cortical malformation. American Journal of Medical Genetics Part A 140: 2416-2425.
190. Sztriha L, Guerrini R, Harding B, Stewart F, Chelloug N, et al. (2004) Clinical, MRI, and pathological features of polymicrogyria in chromosome 22q11 deletion syndrome. American Journal of Medical Genetics Part A 127A: 313-317.
191. Meechan D, Maynard T, Gopalakrishna D, Wu Y (2006) When half is not enough: gene expression and dosage in the 22q11 deletion syndrome. Gene expression 13: 299-310.
192. Gothelf D, Law AJ, Frisch A, Chen J, Zarchi O, et al. (2014) Biological Effects of COMT Haplotypes and Psychosis Risk in 22q11. 2 Deletion Syndrome. Biological psychiatry 75: 406-413.
193. Tsurusaki Y, Koshimizu E, Ohashi H, Phadke S, Kou I, et al. (2014) De novo SOX11 mutations cause Coffin–Siris syndrome. Nature Communications 5.
194. Laumonnier F, Ronce N, Hamel BCJ, Thomas P, Lespinasse J, et al. (2002) Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. American Journal of Human Genetics 71: 1450-1455.

Reference List

195. Numakura C, Kitanaka S, Kato M, Ishikawa S, Hamamoto Y, et al. (2010) Supernumerary impacted teeth in a patient with SOX2 anophthalmia syndrome. American Journal of Medical Genetics Part A 152: 2355-2359.
196. Czakó M, Riegel M, Morava É, Bajnóczky K, Kosztolányi G (2004) Opitz "C" trigonocephaly - like syndrome in a patient with terminal deletion of 2p and partial duplication of 17q. American Journal of Medical Genetics Part A 131: 310-312.
197. Wang Y, Lin L, Lai H, Parada LF, Lei L (2013) Transcription factor Sox11 is essential for both embryonic and adult neurogenesis. Developmental Dynamics 242: 638-653.
198. Wang X, Bjorklund S, Wasik AM, Grandien A, Andersson P, et al. (2010) Gene Expression Profiling and Chromatin Immunoprecipitation Identify DBN1, SETMAR and HIG2 as Direct Targets of SOX11 in Mantle Cell Lymphoma. Plos One 5.
199. Kumar RA, Pilz DT, Babatz TD, Cushion TD, Harvey K, et al. (2010) TUBA1A mutations cause wide spectrum lissencephaly (smooth brain) and suggest that multiple neuronal migration pathways converge on alpha tubulins. Human molecular genetics 19: 2817-2827.
200. Poirier K, Saillour Y, Bahi-Buisson N, Jaglin XH, Fallet-Bianco C, et al. (2010) Mutations in the neuronal ss-tubulin subunit TUBB3 result in malformation of cortical development and neuronal migration defects. Hum Mol Genet 19: 4462-4473.
201. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, et al. (2013) Actionable, pathogenic incidental findings in 1,000 participants' exomes. The American Journal of Human Genetics 93: 631-640.
202. Ming JE, Kaupas ME, Roessler E, Brunner HG, Golabi M, et al. (2002) Mutations in PATCHED-1, the receptor for SONIC HEDGEHOG, are associated with holoprosencephaly. Human genetics 110: 297-301.
203. Quinlan AR, Marth GT (2007) Primer-site SNPs mask mutations. Nature methods 4: 192-192.
204. Landsverk ML, Douglas GV, Tang S, Zhang VW, Wang G-L, et al. (2012) Diagnostic approaches to apparent homozygosity. Genetics in Medicine 14: 877-882.
205. Biesecker LG, Spinner NB (2013) A genomic view of mosaicism and human disease. Nature Reviews Genetics 14: 307-320.
206. Bierhals T, Korenke GC, Uyanik G, Kutsche K (2013) Pontocerebellar hypoplasia type 2 and TSEN2: Review of the literature and two novel mutations. European journal of medical genetics 56: 325-330.
207. Rocas D, Alix E, Michel J, Cordier M-P, Labalme A, et al. (2013) Neuropathological features in a female fetus with OPHN1 deletion and cerebellar hypoplasia. European journal of medical genetics 56: 270-273.
208. Beltrán-Valero de Bernabé D, Currier S, Steinbrecher A, Celli J, van Beusekom E, et al. (2002) Mutations in the O-Mannosyltransferase Gene POMT1 Give Rise to the Severe Neuronal Migration Disorder Walker-Warburg Syndrome. The American Journal of Human Genetics 71: 1033-1043.

Reference List

209. Biesecker LG, Green RC (2014) Diagnostic Clinical Genome and Exome Sequencing. New England Journal of Medicine 370: 2418-2425.
210. Thompson MD, Killoran A, Percy ME, Nezarati M, Cole DE, et al. (2006) Hyperphosphatasia with neurologic deficit: a pyridoxine-responsive seizure disorder? Pediatric neurology 34: 303-307.
211. Board NC (2013) Clinical Commissioning Policy: Pre-implantation Genetic Diagnosis (PGD) http://www.england.nhs.uk/wp-content/uploads/2013/04/e01-p-a.pdf.
212. Schulze TG, McMahon FJ (2005) Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping. Human heredity 58: 131-138.
213. Murakami Y, Tawamie H, Maeda Y, Büttner C, Buchert R, et al. (2014) Null Mutation in PGAP1 Impairing Gpi-Anchor Maturation in Patients with Intellectual Disability and Encephalopathy. PLoS genetics 10: e1004320.
214. Thompson MD, Roscioli T, Marcelis C, Nezarati MM, Stolte-Dijkstra I, et al. (2012) Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). American Journal of Medical Genetics Part A 158A: 553-558.
215. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research 21: 974-984.
216. Ligt J, Boone PM, Pfundt R, Vissers LE, Richmond T, et al. (2013) Detection of Clinically Relevant Copy Number Variants with Whole‐Exome Sequencing. Human mutation 34: 1439-1448.

Appendix

Appendix Table 1: 436 Brain Abnormality Genes

Listed are the 436 brain abnormality genes from the Ingenuity Knowledge base. Key Term used to curate the list was simply "Brain Abnormalities". (Table continues on next page). The entire CCDS annotated coding regions of all 436 genes were targeted by the SeqCap (NimbleGen) capture kit.

ABI1	CASP3	DEAF1	FGF21	HEPACAM	IL8
ACVR1B	CASP7	DHCR7	FGF22	HES1	INPP5E
ACVRL1	CASP9	DICER1	FGF23	HESX1	INTU
ADIPOQ	CAST	DISP1	FGF3	HHAT	IRAK4
AHCY	CC2D2A	DKK1	FGF4	HIPK1	ISPD
AHI1	CCDC88C	DLX5	FGF5	HIPK2	ITGAV
AKT3	CCNA1	DNM1L	FGF6	HMGB1	ITGB2
ALKBH1	CCR2	DRAXIN	FGF7	HMOX1	ITM2B
ALS2	CD93	DVL1	FGF8	HPGDS	JMJD6
ALX1	CDC6	DVL2	FGF9	HSD17B2	KAT2A
ALX4	CDK5RAP2	DYRK1A	FGFR1	HSPG2	KDM6A
APAF1	CDON	E2F5	FIGF	HTT	KIAA1279
APOB	CDT1	E2F6	FKBP1A	IER3IP1	KIF11
APOE	CECR2	EFNA5	FKBP8	IFNG	KIF3B
APP	CENPJ	EFTUD2	FKRP	IFNGR1	KIF7
AQP4	CEP135	EMG1	FKTN	IFT122	KLF2
ARFGEF2	CEP152	EMX2	FLNA	IFT172	KNG1
ARHGAP35	CEP63	ENG	FLT1	IFT57	L1CAM
ARID1B	CERS2	EOMES	FLVCR2	IFT88	LAMC1
ARX	CFL1	EP300	FOXC1	IGBP1	LAMC3
ASPM	CHMP1A	EPHA2	FOXG1	IGF1	LAMP1
ATMIN	CHRD	EPHA7	FOXH1	IL10	LARGE
ATP2C1	CITED2	ERBB3	FOXP2	IL11	LEP
ATR	CLCN7	ERCC6	FURIN	IL12A	LHX2
ATRIP	CLEC1B	ERMARD	GADD45A	IL12B	LIF
B3GNT1	CNR2	EXOSC3	GAS1	IL13	LIG4
B3GNT2	CNTF	EYA1	GDF1	IL17A	LMNB1
BAG6	CNTNAP2	FAM20C	GHRHR	IL17B	LMNB2
BCL10	COL18A1	FGF1	GJB2	IL17C	LMO4
BCL2L1	COL2A1	FGF10	GLDC	IL17D	LPAR1
BMI1	COX7B	FGF11	GLI2	IL17F	LRP2
BMP7	CREBBP	FGF12	GMPPB	IL18	LRP5
BMPER	CSF2	FGF13	GPR161	IL1A	LRP6
BMPR1A	CSF3	FGF14	GPR56	IL1B	LTA
BRCA1	CSTB	FGF16	GRHL3	IL2	MAOB
BRCA2	CTNNBIP1	FGF17	GRK5	IL25	MAP1B
CALR	CYP26A1	FGF18	GSK3B	IL3	MAP3K4
CARD10	CYP26C1	FGF19	HDAC4	IL4	MAPK8
CASC5	DCC	FGF2	HDAC8	IL5	MAPK9
CASK	DCX	FGF20	HECTD1	IL6	MAPT

Appendix Table 1: 436 Brain Abnormality Genes (continued from previous page)

MARCKS	PAFAH1B1	PTGDR	SLC16A2	TP53
MARCKSL1	PALB2	PTGES	SLC25A19	TRAF6
MCPH1	Palld	PTPN9	SMAD2	TRPM6
MDM2	PARP1	PTPRF	SMAD3	TSC1
MDM4	PATZ1	PTPRJ	SMAD5	TSC2
MECP2	PAX2	PTPRS	SMARCA4	TSEN2
MED17	PAX6	PURA	SMARCB1	TSEN34
MEF2C	PCNT	PYGO1	SMARCC1	TSEN54
MEN1	PDGFB	PYGO2	SMO	TUBA1A
MESP1	PDGFC	QPCT	SMURF1	TUBA8
MIB2	PDS5B	RAB3GAP1	SMURF2	TUBB2B
MIR17HG	PEX2	RALA	SNX1	TUBB3
MLC1	PGF	RALB	SNX2	TUBGCP6
MPDZ	PHF6	RAPGEF2	SOD1	TWIST1
MRPS16	PHGDH	RARS2	SOD2	TWSG1
MTHFD1L	PIGA	RBBP8	SOD3	UBE3B
MTHFR	PIK3CA	RBPJ	SP8	VEGFA
MYCN	PIK3R2	RELN	SPHK1	VEGFB
NBN	PITX2	RFX4	SPHK2	VEGFC
NDE1	PLAT	RGMA	SPRY2	VLDLR
NDST1	PLAU	ROR1	SPRY4	VPS26A
NDST3	PLXNB2	ROR2	SRPX2	VRK1
NDUFS4	PNPO	RPGRIP1L	STAMBP	WDR62
NEUROD1	POMGNT1	RPS6KA3	STIL	XPA
NF1	POMGNT2	RPTOR	TCOF1	XRCC2
NFIB	POMK	RTTN	TCTN2	YBX1
NFIX	POMT1	S100B	TDGF1	YBX3
NHEJ1	POMT2	SALL1	TEAD1	YWHAE
NIN	POR	SALL2	TEAD2	YY1
NODAL	PRKACA	SALL4	TERT	ZBTB16
NOG	PRKACB	SCAP	TFAP2A	ZIC1
NQO1	PRNP	SCN2A	TFDP1	ZIC2
NRXN1	PROC	SEPSECS	TGFB1	ZIC4
NUAK2	PROCR	SERPINE1	TGFB2	ZIC5
OCLN	PROK1	SERPINI1	TGFB3	ZNF335
ORC1	PROS1	SHH	TGIF1	ZNF423
ORC4	PRRX1	SIRT1	THOC6	
ORC6	PSEN1	SIX3	TMEM5	
OSM	PSEN2	SKI	TMEM67	
OTX2	PTCH1	SLC12A6	TNF	

Appendix Table 2: Genes of interest that are not completely targeted by the Illumina TruSeq capture kit.

Of the 436 genes of interest (Appendix Table 1), 68 are not completely targeted by the TruSeq capture kit based on the CCDS exonic regions. 2 genes, *CNTF* and *GDF1* are not targeted at all. The kit targets in total 98.71% of all positions in the genes of interest.

Gene	Percent targeted	Gene	Percent targeted
AHI1	97.13%	IFT88	99.16%
AKT3	97.78%	KLF2	23.50%
APP	91.22%	LRP5	98.05%
ARID1B	99.67%	MAOB	97.06%
ATP2C1	96.63%	MAP3K4	96.82%
CARD10	93.22%	MCPH1	96.89%
CASC5	96.84%	MIB2	98.63%
CAST	95.00%	MYCN	99.84%
CC2D2A	99.08%	NDE1	93.95%
CCDC88C	99.01%	NF1	99.29%
CDK5RAP2	97.69%	NIN	96.95%
CHRD	94.84%	OCLN	97.32%
CLEC1B	80.33%	ORC4	92.99%
CNTF	0.00%	PATZ1	96.85%
COL18A1	93.32%	PIK3R2	96.71%
COL2A1	96.34%	PLAU	80.09%
CREBBP	96.63%	POMT1	97.06%
CYP26C1	78.90%	PRKACA	95.64%
DEAF1	87.05%	RPGRIP1L	96.50%
E2F5	77.52%	RPGRIP1L	99.12%
ENG	98.62%	RTTN	97.83%
ERCC6	94.47%	SALL1	99.90%
EYA1	99.16%	SCN2A	95.21%
FGF16	99.43%	SMARCC1	94.12%
FGF22	99.81%	SOD2	95.83%
FGF8	89.35%	TCOF1	97.69%
FGFR1	96.35%	TCTN2	94.75%
FKBP8	76.49%	TDGF1	80.60%
FOXP2	90.64%	TSEN54	96.46%
GDF1	0.00%	TUBA8	74.95%
GRHL3	97.95%	VEGFA	95.88%
GRK5	97.07%	VEGFB	90.38%
HDAC4	95.24%	VLDLR	96.77%
HSPG2	95.07%	YBX1	82.97%

REPORT

Mutations in *PGAP3* Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation

Malcolm F. Howard,[1,13] Yoshiko Murakami,[2,3,13] Alistair T. Pagnamenta,[1,13] Cornelia Daumer-Haas,[4] Björn Fischer,[5,7] Jochen Hecht,[6,7] David A. Keays,[8] Samantha J.L. Knight,[1] Uwe Kölsch,[9] Ulrike Krüger,[5] Steffen Leiz,[10] Yusuke Maeda,[2,3] Daphne Mitchell,[11] Stefan Mundlos,[5,6,7] John A. Phillips III,[11] Peter N. Robinson,[5,6,7] Usha Kini,[12,14,*] Jenny C. Taylor,[1,14] Denise Horn,[5,14] Taroh Kinoshita,[2,3,14] and Peter M. Krawitz[5,6,7,14,*]

Glycosylphophatidylinositol (GPI)-anchored proteins play important roles in many biological processes, and mutations affecting proteins involved in the synthesis of the GPI anchor are reported to cause a wide spectrum of intellectual disabilities (IDs) with characteristic additional phenotypic features. Here, we describe a total of five individuals (from three unrelated families) in whom we identified mutations in *PGAP3*, encoding a protein that is involved in GPI-anchor maturation. Three siblings in a consanguineous Pakistani family presented with profound developmental delay, severe ID, no speech, psychomotor delay, and postnatal microcephaly. A combination of autozygosity mapping and exome sequencing identified a 13.8 Mb region harboring a homozygous c.275G>A (p.Gly92Asp) variant in *PGAP3* region 17q11.2–q21.32. Subsequent testing showed elevated serum alkaline phosphatase (ALP), a GPI-anchored enzyme, in all three affected children. In two unrelated individuals in a cohort with developmental delay, ID, and elevated ALP, we identified compound-heterozygous variants c.439dupC (p.Leu147Profs*16) and c.914A>G (p.Asp305Gly) and homozygous variant c.314C>G (p.Pro105Arg). The 1 bp duplication causes a frameshift and nonsense-mediated decay. Further evidence supporting pathogenicity of the missense mutations c.275G>A, c.314C>G, and c.914A>G was provided by the absence of the variants from ethnically matched controls, phylogenetic conservation, and functional studies on Chinese hamster ovary cell lines. Taken together with recent data on *PGAP2*, these results confirm the importance of the later GPI-anchor remodelling steps for normal neuronal development. Impairment of PGAP3 causes a subtype of hyperphosphatasia with ID, a congenital disorder of glycosylation that is also referred to as Mabry syndrome.

Glycosylphophatidylinositol (GPI) anchoring is a post-translational modification that tethers proteins to plasma membranes, and it is thought to play a role in protein sorting and trafficking.[1] The GPI anchor is well conserved among eukaryotes, and there are over 150 mammalian GPI-anchored proteins (GPI-APs), including receptors, adhesion molecules, and enzymes.[1] Many of these proteins are critical for normal neural and embryonic development.[2,3] There are around 30 genes known to be involved in the biosynthesis and remodelling of the GPI anchor, which is formed in the endoplasmic reticulum (ER), where it is attached by the GPI transamidase to a protein showing a specific C-terminal signal before it is transported to the Golgi apparatus for fatty acid remodelling and cellular export.

In the last 4 years, germline mutations in eight genes selectively involved in the GPI-anchor-synthesis pathway have been shown to cause a wide phenotypic spectrum of disorders with intellectual disability (ID) and seizures; these range from syndromic forms with characteristic malformations and minor anomalies to nonsyndromic forms. Congenital disorders that are caused by an impairment of GPI-anchor synthesis and maturation are now classified as congenital disorders of glycosylation (CDGs), a diverse class of metabolic diseases.[4]

PIGV (MIM 610274), *PIGO* (MIM 614730), and *PGAP2* (MIM 615187), three different genes in the GPI-anchor-synthesis pathway, have been implicated in hyperphosphatasia with mental retardation syndrome (HPMRS [MIM 239300]), also known as Mabry syndrome, a recently delineated autosomal-recessive form of ID with a distinct facial dysmorphism, consistently elevated serum alkaline phosphatase (ALP) (hyperphosphatasia), brachytelephalangy, and seizures. All individuals with congenital impairment of GPI-anchor synthesis have residual surface levels of GPI-anchored proteins, and it is hypothesized that complete-loss-of-function mutations in genes involved with GPI-anchor synthesis are embryonically lethal.[5]

[1]National Institute for Health Research Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN Oxford, UK; [2]Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, Japan; [3]World Premier International Immunology Frontier Research Center, Osaka University, Osaka 565-0871, Japan; [4]Pränatal-Medizin München, 80637 München, Germany; [5]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, 13353 Berlin, Germany; [6]Berlin Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin, 13353 Berlin, Germany; [7]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [8]Research Institute of Molecular Pathology, 1030 Vienna, Austria; [9]Institute of Medical Immunology, Charité Universitätsmedizin, 13353 Berlin, Germany; [10]Department of Pediatrics, Klinikum Dritter Orden, 80638 München, Germany; [11]Division of Medical Genetics and Genomic Medicine, Department of Pediatrics, Vanderbilt University School of Medicine, TN 37232-2578, USA; [12]Department of Clinical Genetics, Oxford University Hospitals NHS Trust, OX3 9DU Oxford, UK
[13]These authors contributed equally to this work
[14]These authors contributed equally to this work and are co-senior authors
*Correspondence: usha.kini@ouh.nhs.uk (U.K.), peter.krawitz@charite.de (P.M.K.)
http://dx.doi.org/10.1016/j.ajhg.2013.12.012. ©2014 by The American Society of Human Genetics. All rights reserved.

The effect of mutations causing GPI-anchor deficiencies can be subdivided into groups on the basis of the serum activity of the marker enzyme ALP. Mutations in *PIGA* (MIM 311770), *PIGL* (MIM 605947), *PIGM* (MIM 610273), *PIGN* (MIM 606097), and *PIGT* (MIM 610272) affect early GPI-anchor synthesis and the attachment of proteins to the GPI anchor.[6–10] These variants result in primary reduced surface levels of GPI-APs because of their increased intracellular degradation and are not associated with high serum ALP. Mutations in *PIGV* and *PIGO* affect late GPI-anchor synthesis.[11,12] However, GPI transamidase recognizes the incomplete GPI anchor and cleaves the GPI attachment signal, leading to the generation of non-GPI-anchored soluble proteins.[13] Individuals with mutations in *PIGV* and *PIGO* have primary reduced GPI-AP surface levels as a result of increased secretion into the extracellular space, resulting in high serum ALP. Recently, mutations in *PGAP2* have been reported to affect the final GPI-anchor fatty-acid-remodelling step and to thus result in abnormal GPI-APs that are more prone to cleavage.[14–16] Individuals with *PGAP2* mutations have secondary reduced GPI-AP surface levels also as a result of increased secretion into the extracellular space, resulting in high serum ALP. GPI remodelling occurs in the Golgi and involves the removal of an unsaturated fatty acid at the sn-2 position by PGAP3 and its subsequent substitution with a saturated fatty acid by PGAP2. Fatty acid remodelling is critical for proper association between GPI-APs and lipid rafts.[17]

A number of deficient GPI-pathway genes were first identified via traditional autozygosity mapping in combination with next-generation sequencing methodologies.[10,14] This methodology has become a key strategy for the identification of disease-associated genes, particularly those related to genetically heterogeneous conditions such as ID, and was the approach taken for the first family described here.[18]

Family A originates from the Sargodha district of Punjab in Pakistan. The parents are second cousins, and all three children (Figures 1A and 2A) presented with profound developmental delay, severe learning disability, no speech, psychomotor delay, and postnatal microcephaly (−2 to −3 SDs) at the ages of 17, 8, and 4 years. Further clinical details are summarized in Table 1. After a *ASPM* test that did not detect pathogenic variants, they were recruited into the ongoing Structural Brain Abnormalities and Learning Disabilities Study (see Oxford Brain Abnormality Research Group in Web Resources), which received UK ethics approval from the Wales Research Ethics Committee (12/WA/0001) and obtained informed consent from the responsible persons on behalf of all study participants. Blood DNA samples for IV-2, IV-3, V-1, V-2, and V-3 were run on CytoSNP12v2 arrays (Illumina), which confirmed familial relationships and that there were no copy-number changes of likely clinical relevance (Table S1, available online). Autozygosity mapping was then used for identifying a single 13.8 Mb candidate region in 17q11.2–q21.32 (Figure S1). None of the 359 annotated genes appeared to be obvious candidates, and so an exome sequencing approach was adopted.

The exome of proband V-2 was enriched (TruSeq, Illumina) and sequenced on a HiSeq2500 (Illumina) with the use of standard settings and the 100 bp paired-end read format. Reads were mapped to the human reference sequence (UCSC Genome Browser, hg19) with the use of Stampy,[19] and variants were called with Platypus. The resulting VCF file was uploaded into Ingenuity Variant Analysis (v.2.1.20130711) and filtered for rare variants that were in the homozygous region on chromosome 17 and predicted to be deleterious (Table S2 and Web Resources).

The single remaining variant was a c.275G>A change in *PGAP3* (RefSeq accession number NM_033419.3). Sanger sequencing confirmed that this variant cosegregated with disease, consistent with an autosomal-recessive mode of inheritance (Figure 1B). The variant is predicted to cause a p.Gly92Asp alteration to the protein sequence at a highly conserved position (Figure 1C) and is not present at any frequency in the NHLBI Exome Sequencing Project Exome Variant Server (EVS), 1000 Genomes (version 3), or 274 in-house genomes of mixed ancestry. The mutation was also not detected in 108 Punjabi individuals from Lahore, suggesting that c.275G>A is unlikely to be a variant specific to this ethnicity.

In parallel, we performed targeted sequencing in 19 individuals with ID and elevated ALP and in whom HPMRS was suspected. The Charité University Medicine ethics board approved this study, and we obtained informed consent from the responsible persons (parents) on behalf of all study participants. After excluding pathogenic *PIGV* mutations that are the most common cause of HPMRS by Sanger sequencing, we subjected DNA to enrichment of all 30 known genes associated with GPI-anchor synthesis by using a customized SureSelect library (Agilent), as well as subsequent sequencing on a HiSeq2000, as previously described.[15] Sequence variants were filtered under an autosomal-recessive model of inheritance in GeneTalk,[20] which yielded potentially pathogenic variants in *PGAP3* in two individuals. Compound-heterozygous variants c.439dupC (p.Leu147Profs*16) and c.914A>G (p.Asp305Gly) were detected in individual II-1 from family B, and homozygous variant c.314C>G (p.Pro105Arg) was identified in individual II-1 from family C (Figure 1B).

The detailed clinical findings of individual II-1 from family B were reported previously (affected individual 4 in Thompson et al.)[21] together with a single rare heterozygous variant in *PIGV*, c.1369C>T.[21] However, in *PIGV*-defective Chinese hamster ovary (CHO) cells, the altered protein restored the surface levels of GPI-APs as efficiently as the wild-type protein, suggesting no functional impairment of *PIGV* (Figure S4). In contrast, the duplication of c.439dupC in *PGAP3* causes a frameshift that introduces a premature stop codon, most likely resulting in nonsense-mediated decay of the transcript. To exclude further candidate mutations, we exome sequenced DNAs of the affected

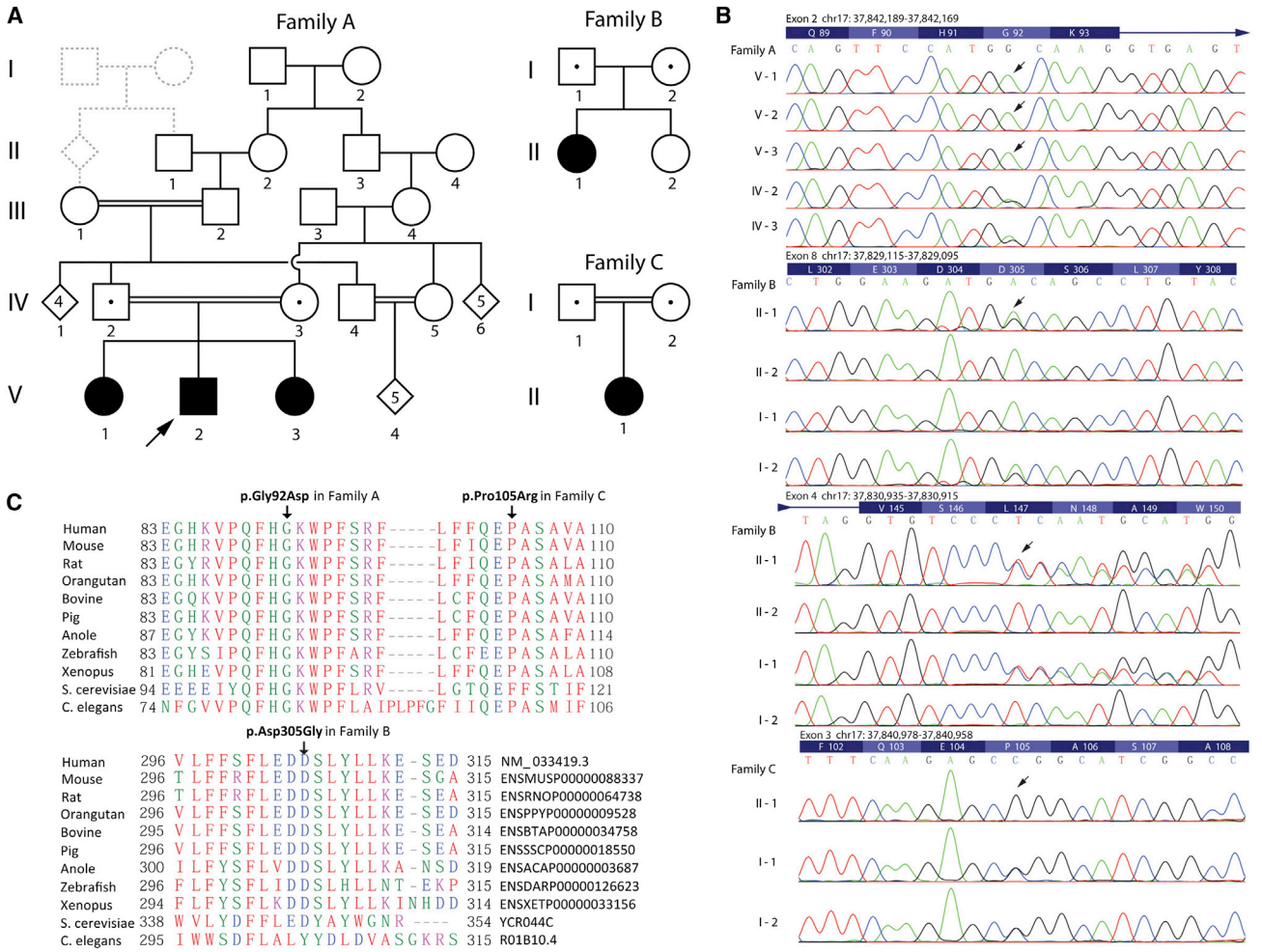


Figure 1. Pedigrees, Sanger Validation, and Phylogenetic Conservation of the Variant Filtering Cascade and Molecular Characterization of *PGAP3* Mutations

(A) Pedigrees are shown for family A (of Pakistani descent), family B (of European descent), and family C (of Saudi Arabian origin). Shading indicates ID and hyperphosphatasia, and proven heterozygote carriers are shown by a dot. In family A, the gray dotted lines indicate an additional consanguineous loop that was not described in the clinical notes but that was inferred by the high inbreeding coefficient detected in IV-2 from the SNP array data. The proband is indicated with an arrow.

(B) Sanger validation and segregation testing of *PGAP3* variants. DNA for the unaffected double first cousins in family A (collectively labeled as V-4) was not available.

(C) ClustalW alignment of amino acid sequence shows high evolutionary conservation.

individual of family B and her parents and confirmed the compound-heterozygous variants in *PGAP3* as the only candidate mutations for an autosomal-recessive model of inheritance.[22,23]

Individual II-1 of family B is the first child of nonconsanguineous European-American parents (Table 1 and Figures 1A and 2B). Her psychomotor development was severely delayed, and she developed tonic-clonic seizures. Physical examination at the age of 10 years showed postnatal short stature, dystrophy, normal occipitofrontal circumference (OFC), and dysmorphic facial features. ALP levels were repeatedly elevated, and a GPI-anchor deficiency on granulocytes and monocytes could be measured by flow cytometry with CD16, CD24, and fluorescence-labeled aerolysin staining (Figure 3A). In family C, individual II-1 is the first child of consanguineous parents of Saudi Arabian origin (Table 1 and Figures 1A and 2C). She had severe psychomotor delay. Myoclonic seizures started in her second year of life. Physical examination of this 2-year-old female showed normal growth parameters and OFC but axial muscular hypotonia, uncoordinated movements, and facial dysmorphism. ALP activity was elevated in repeated tests. Sanger sequencing confirmed that the parents are heterozygous for c.314C>G, and the mutation was not detected in 52 Arabic controls or the NHLBI Exome Sequencing Project EVS, suggesting that it might be a private mutation specific to this family (Figure 1B).

The most common features noted in all affected individuals were normal growth parameters and OFC at birth, severe psychomotor delay with no speech and marked motor delay, and characteristic facial features. Four were unable to walk, whereas the other began to walk with support at the age of 3. Seizures starting between the ages of 18 months and 12 years developed in four of five affected



Figure 2. Photographs of Affected Individuals
Facial features seen in (A) individual V-3 from family A, (B) individual II-1 from family B, and (C) individual II-1 from family C at the ages of 4, 10, and 2 years, respectively. These individuals bear a striking resemblance with a broad nasal bridge, long-appearing palpebral fissures, a broad nasal tip, a short nose, a long philtrum, a thin and wide upper lip, full cheeks, and large fleshy ear lobes.

individuals. Three were described as having tonic-clonic seizures, and the other had myoclonic seizures. Consistent ALP elevation varied between 1.4 and 5 times the age-adjusted upper limit of the normal range. The facial gestalt was similar to that of individuals with HPMRS and included features such as a broad nasal bridge, a short nose with a broad nasal tip, a thin and wide upper lip, and large fleshy ear lobes. In contrast, postnatal development of head circumference and growth parameters differed among the affected individuals studied here. Affected individuals in family A developed microcephaly, whereas OFC was in the mean range in the other individuals. In the affected individual from family B, marked short stature was present, whereas the other affected individuals had a mean height according to age. Two of five individuals had cleft palate, whereas abnormal MRI findings such as thin corpus callosum and dilated lateral ventricles were observed in the affected individual from family C.

The common features of all individuals reported with mutations in *PIGV*, *PIGO*, *PGAP2*, and *PGAP3* are severe psychomotor delay and ID, epilepsy, elevated ALP, and a distinctive facial gestalt. In contrast, brachytelephalangy, which is an important diagnostic sign in other types of HPMRS, is not present in any affected individuals with *PGAP3* mutations. Marked variability regarding the postnatal growth and OFC has also been observed in the groups of affected individuals carrying *PIGV*, *PIGO*, and *PGAP2* mutations. Also, cleft palate has been previously documented in other individuals with HPMRS due to mutations in these genes. In individuals affected by HPMRS and mutations in genes other than *PGAP3*, associated malformations seem to appear more frequently and their spectrum seems to be broader.

To determine the functional consequences of the four *PGAP3* mutations, we used a mutant CHO cell line defective in both *PGAP3* and *PGAP2* (see Figure S3 for principles of the assay).[17] The mutant cells have GPI-APs at mildly reduced levels because of a lack of GPI fatty acid remodelling. When wild-type *PGAP3* cDNA was transfected, the first step in the fatty acid remodelling was restored, whereas the second step remained defective, leading to the release of lyso-GPI intermediates and resulting in a severe reduction in the surface levels of GPI-APs (Figure 3Bi). If the mutant *PGAP3* cDNA has decreased activity, reduction in the surface GPI-AP levels would be partial. Mutant *PGAP3* cDNA bearing the mutation found in family A (c.275G>A [p.Gly92Asp]) either did not reduce or reduced only slightly the surface levels of three GPI-APs: CD59, CD55 (DAF), and urokinase plasminogen activator receptor (uPAR), indicating that the substitution caused a null or nearly null phenotype (Figure 3Bii). One of the substitutions found in family B (p.Leu147Profs*16) also caused a null phenotype, whereas the other (p.Asp305Gly) significantly reduced levels of all three GPI-APs, indicating some residual activity (Figure 3Biii). The substitution found in family C (p.Pro105Arg) caused less efficient reduction of GPI-APs than did the p.Asp305Gly substitution in family B, indicating an even lower residual activity (Figure 3Biv).

To clarify the mechanisms of these functional losses, we expressed mutant *PGAP3* clones that were hemagglutinin (HA) tagged at the N terminus in CHO cells and analyzed them by SDS-PAGE and immunoblotting (Figure 4) and immunofluorescence microscopy (Figure 5) with anti-HA. On SDS-PAGE and immunoblotting, HA-PGAP3 appeared as a smear band at around 37–45 kDa and several clear bands at around 23–35 kDa (Figure 4A, lane 3). After treatment with PNGase F to remove N-glycan, the smear band disappeared and a band at 33 kDa corresponding to de-N-glycosylated full-size protein became a major one (lane 1). The altered (p.Gly92Asp) protein in family A showed similar profiles (lanes 7 and 9), suggesting normal levels of PGAP3 bearing normally matured N-glycan. Consistently, on immunofluorescence microscopy,

Table 1. Summary of Clinical Findings in Individuals with *PGAP3* Mutations

	Individual				
Clinical Findings[a]	**V-1 (Family A)**	**V-2 (Family A)**	**V-3 (Family A)**	**II-1 (Family B)**	**II-1 (Family C)**
Ethnicity	Pakistani (Sargodha district of Punjab)	Pakistani (Sargodha district of Punjab)	Pakistani (Sargodha district of Punjab)	American (European descent)	Saudi-Arabian
Consanguinity	yes	yes	yes	no	yes
Age of last assessment (years)	17	8	4	10	2
OFC at birth	normal	normal	normal	normal	normal
OFC	−2 to −3 SDs	−2 to −3 SDs	−3 SDs	mean	mean
Height	normal	normal	normal	−4 SDs	normal
Weight	normal	normal	normal	−2 SDs	normal
Global developmental delay (HP:0001263)	yes	yes	yes	yes	yes
Motor delay (HP:0001270)	severe (unable to walk)	severe (unable to walk)	severe (unable to walk)	severe (walk with support at age of 3 years)	severe (unable to walk)
Speech and language development	none	none	none	none	none
Muscular hypotonia (HP:0001252)	yes	yes	yes	yes	yes
Seizures (HP:0001250)	yes	yes	no	yes	yes
Age of onset of seizures (years)	12	4	NA	4	1.5
Type of seizures	generalized tonic-clonic	generalized tonic-clonic	NA	tonic-clonic and cluster	myoclonic
Antiepileptic drugs	valproate	valproate	NA	rufinamide, pregabaline	valproate, levetiracetam
Behavioral abnormalities	involuntary midline hand movements, bruxism	involuntary midline hand movements, bruxism	involuntary midline hand movements, bruxism	no	hyperactivity
Apparent hypertelorism (HP:0000316)	yes	yes	yes	yes	yes
Upslanting palpebral fissures (HP:0000582)	yes	yes	yes	no	no
Broad nasal bridge (HP:0000431)	yes	yes	yes	yes	yes
Broad nasal tip	yes	yes	yes	yes	yes
Short nose (HP:0003196)	yes	yes	yes	yes	yes
Tented upper-lip vermilion (HP:0010804)	yes	yes	yes	yes	yes
Large, fleshy ear lobes (HP:0009748)	yes	yes	yes	no	yes
Cleft palate (HP:0000175)	yes	yes	no (high palate)	no	no
Brachytelephalangy (HP:0009882)	no	no	no	no	no
Serum total ALP (U/l) (HP:0003155)	739	1,407	1,452	926–2,000	928–1,370
Upper limit in ALP test (U/l)	525–600	525–600	525–600	400	386

(*Continued on next page*)

Table 1. ***Continued***

Clinical Findings[a]	Individual				
	V-1 (Family A)	V-2 (Family A)	V-3 (Family A)	II-1 (Family B)	II-1 (Family C)
Further anomalies	no	no	no	no	thin corpus callosum, dilated lateral ventricles
PGAP3 variants (RefSeq NM_033419.3)	homozygous c.275G>A (p.Gly92Asp)	homozygous c.275G>A (p.Gly92Asp)	homozygous c.275G>A p.Gly92Asp)	compound-heterozygous c.914A>G (p.Asp305Gly) and c.439dupC (p.Leu147Profs*16)	homozygous c.314C>G (p.Pro105Arg)

Abbreviations are as follows: ALP, alkaline phosphatase; OFC, occipitofrontal circumference; and NA, not applicable.
[a]Human phenotype ontology IDs are provided if applicable.

wild-type and p.Gly92Asp PGAP3 were found mainly in the Golgi and to a lesser extent in the ER (Figure 5, first and second rows). Only faint bands were seen with p.Leu147Profs*16 from family B, indicating that the level of altered protein was not significant (Figure 4C). The other family B substitution, p.Asp305Gly, showed a main band at 35 kDa (Figure 4A, lane 6) that was shifted to the 33 kDa de-N-glycosylated full-size protein after treatment with either PNGase F or Endo H, the latter of which eliminated only immature N-glycan (lanes 4 and 5). This profile indicated that the p.Asp305Gly altered protein had only immature ER-form N-glycan. Consistently, the altered

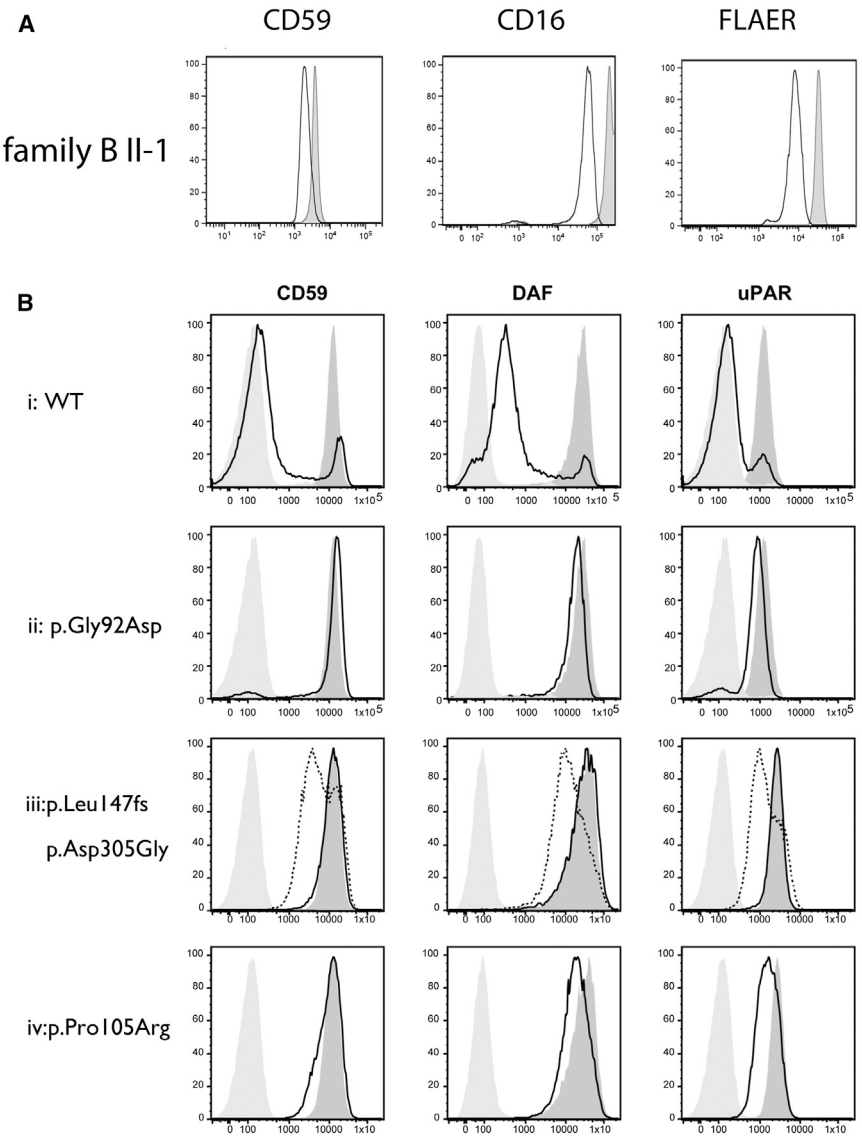


Figure 3. Flow Cytometric Analysis of Granulocyte Surface GPI-APs and Flow Cytometric Assay for Functions of Altered PGAP3

(A) Blood granulocytes from the affected individual in family B (solid line) were stained with anti-CD24 and anti-CD16 and fluorescence-labeled aerolysin (FLAER). The light-gray area represents a healthy control.

(B) *PGAP3* and *PGAP2* double-mutant CHO cells were transfected with *PGAP3* cDNA and 2 days later were stained for CD59, CD55 (DAF), and uPAR. (Bi) After transfection with wild-type *PGAP3* cDNA, the surface levels of three GPI-APs were severely reduced (solid lines). The dark area shows the original surface levels of GPI-APs on *PGAP3* and *PGAP2* double-mutant CHO cells, whereas the light gray area represents the isotype-matched control. (Bii) The *PGAP3* cDNA bearing mutation c.275G>A (p.Gly92Asp) in family A either did not reduce or only slightly reduced GPI-AP levels (solid lines). (Biii) The *PGAP3* cDNA bearing one mutation, c.439dupC (p.Leu147Profs*16), in family B did not reduce GPI-AP levels (solid lines), whereas that bearing the other mutation, c.914A>G (p.Asp305Gly), significantly reduced GPI-AP levels, indicating a hypomorphic mutant phenotype (dotted lines). (Biv) The *PGAP3* cDNA bearing mutation c.314C>G (p.Pro105Arg) in family C slightly reduced the levels of three GPI-APs, indicating a hypomorphic but very severe loss-of-function phenotype (solid lines).

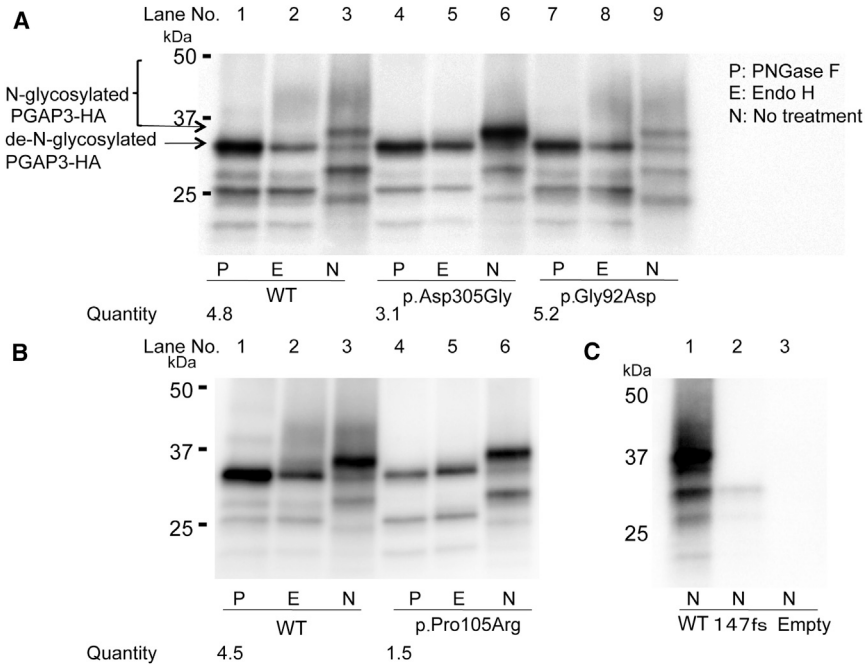


Figure 4. SDS-PAGE and Immunoblotting of HA-Tagged PGAP3

(A) Lanes 1–3, wild-type; lanes 4–6, p.Asp305Gly; lanes 7–9, p.Gly92Asp. Protein levels shown as quantity were normalized with the intensities of GAPDH for the loading control and luciferase activities for the transfection efficiencies. The antibody showed no reactivity in lysates from empty-vector-transfected cells.

(B) Lanes 1–3, wild-type; lanes 4–6, p.Pro105Arg.

(C) Lane 1, wild-type; lane 2, p.Leu147Profs*16; lane 3, empty vector. There are two N-glycosylation sites in PGAP3. The mature PGAP3 appears as a smear at 37–45 KDa (A, lane 3). This smear is sensitive to PNGase F, suggesting heterogeneous N-glycans. Bands seen below the 35 KDa position represent degradation products because they are not seen in the empty-vector transfectant (C, lane 3) or the transfectant with the truncated p.Leu147Profs*16 protein (C, lane 2). Abbreviations are as follows: N, no treatment; P, PNGase F treatment; and E, Endo H treatment.

protein was found in the ER, but not the Golgi (Figure 5, third row). Similarly, the p.Pro105Arg substitution in family C generated only an immature protein with ER-form N-glycan, as shown by a lack of a smear band at around 37–45 kDa (Figure 4B, lane 6) and a shift of the clear major band at 35 kDa to 33 kDa after Endo-H treatment (lanes 5 and 6). Consistent with these results, p.Pro105Arg PGAP3 was localized to the ER, but not to the Golgi (Figure 5, fourth row).

Therefore, the four substitutions that we describe had different effects on PGAP3, the Golgi-resident GPI-specific phospholipase A_2 consisting of an N-terminal luminal domain, seven transmembrane domains with conserved catalytic amino acids, and a C-terminal cytoplasmic tail (Figure S4). (1) The p.Gly92Asp substitution did not affect protein levels or Golgi localization, yet the altered protein had only negligible activity. Because Gly92 resides in a juxta membrane position on the luminal side, a negative charge caused by this substitution might interfere with the association between PGAP3 and GPI anchor substrate. (2) The p.Leu147Profs*16 protein was not detected significantly, consistent with nonsense-mediated mRNA degradation. (3) The p.Asp305Gly and p.Pro105Arg proteins were readily detectable but had immature N-glycan and were mislocalized in the ER. Asp305 resides in the cytoplasmic tail, whereas Pro105 resides in the first transmembrane domain. The cytoplasmic tail and the first transmembrane domain might be important for Golgi localization.

The characteristic biochemical phenotype of individuals with *PGAP3* deficiency is hyperphosphatasia, a sign of GPI-AP release from the cell surface. However, the reduction of GPI-AP levels is expected to depend on the cell type and species, as studies in *Pgap3*-knockout mice suggest.[24,25] In fact, flow cytometric analysis from the affected individual in family B demonstrated significant reduction in the cell-surface levels of GPI-APs in blood granulocytes (Figure 3A), whereas the CD59 surface levels in erythrocytes of an affected individual in family A were found to be normal. This confirms that GPI remodelling by PGAP3 is not essential for CD59 surface levels in erythrocytes, given that the GPI-anchor structure in erythrocytes maintains an unsaturated fatty acid at the sn2 position.[26]

The exact mechanism of release has yet to be characterized, but GPI-APs bearing unremodelled fatty acids are not associated well with lipid rafts.[17] This abnormal membrane distribution might affect the stability of GPI-APs, resulting in release from the cell surface. Although both PGAP3 and PGAP2 are involved in fatty acid remodeling, the above mechanism of hyperphosphatasia or GPI-AP release from *PGAP3*-deficient cells is different from that observed in *PGAP2*-deficient cells. In the latter cells, only the first reaction by PGAP3, elimination of unsaturated fatty acid, occurs and the resulting lyso-GPI intermediate is cleaved and released by a putative lyso-GPI-specific phospholipase D.[15]

In summary, we identified four different *PGAP3* mutations in three unrelated families by using two independent strategies. In family A, recruited on account of postnatal microcephaly, a combination of autozygosity mapping and exome sequencing identified a missense variant in *PGAP3* as the only likely candidate. In families B and C, mutations were uncovered in the same gene via a targeted sequencing approach in a cohort of individuals ascertained specifically for ID and hyperphosphatasia. In spite of the different approaches, the convergent findings, taken together with the segregation, the absence of these

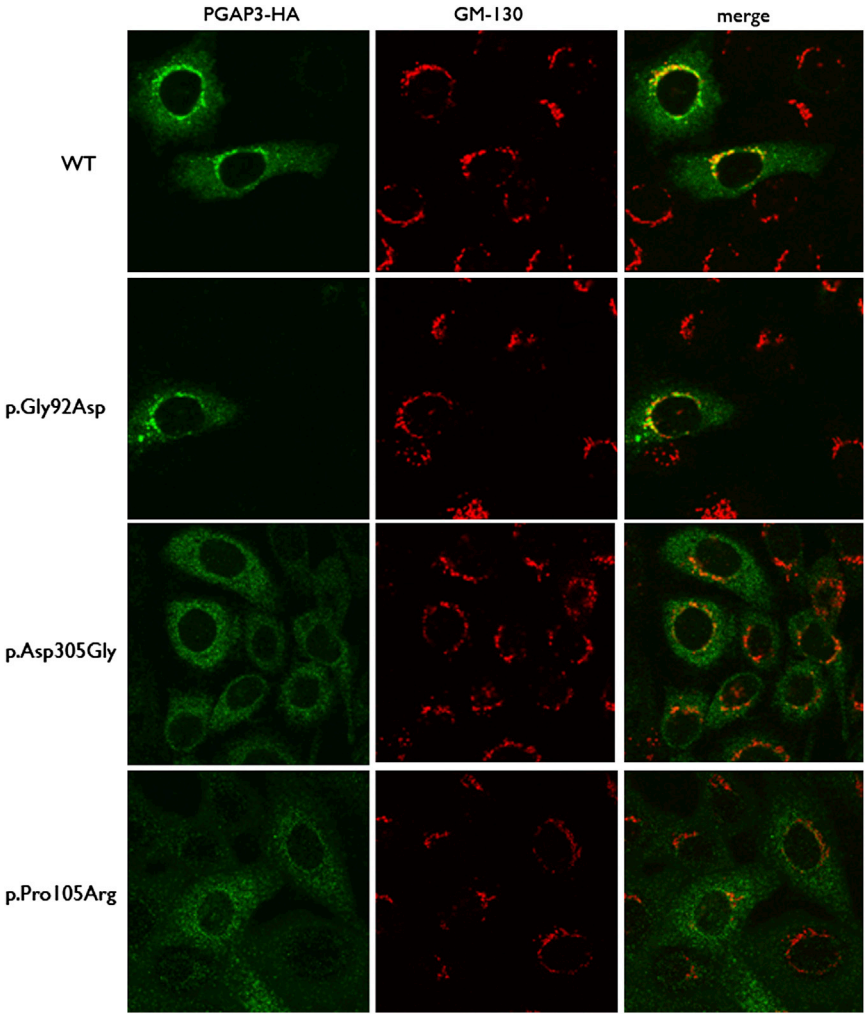


Figure 5. Subcellular Localization of HA-Tagged PGAP3 in CHO Cells
(First row) Wild-type.
(Second row) p.Gly92Asp in family A.
(Third row) p.Asp305Gly in family B.
(Fourth row) Pro105Arg in family C.
GM-130 is the Golgi marker. The relatively high levels of the ER form of HA-tagged wild-type PGAP3 seen in the top panels and also in lane 3 in Figure 4A might have been due to overexpression.

variants in suitable controls, phylogenetic conservation, and functional studies on CHO cell lines, provide strong evidence supporting the etiological role of these mutations. Along with *PGAP1* and *PGAP2*, *PGAP3* is responsible for the modification of the fatty acid residues on the GPI anchor in a maturation process that occurs in the ER and Golgi. Our results and previously reported data on *PGAP2* mutations suggest that impairment of fatty acid remodelling results in GPI-APs that are more prone to cleavage on the plasma membrane and clinical features that are similar to HPMRS. Our findings widen both the phenotypic and the genotypic spectra of hyperphosphatasia and ID syndromes that are caused by mutations in *PIGV, PIGO*, and *PGAP2* and suggest a heterogeneous etiology caused by impairment of late GPI-anchor synthesis and the GPI-AP-maturation pathway. These functional mutations define this condition as a CDG, PGAP3-CDG. In contrast to classical CDGs, which affect N-glycosylation and O-glycosylation, or both, defects in the GPI-anchor-biosynthesis pathway cannot be detected with a transferrin or APOCIII glycosylation assay. This underlines the importance of comprehensive genetic testing of individuals with suspected CDGs.

Supplemental Data

Supplemental Data include four figures and two tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre Oxford with funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health. The work was also supported by a grant from the German Ministry of Research and Education to S.M. (0313911), by Deutsche Forschungsgemeinschaft grants to P.M.K. (KR 3985/1-1) and S.M. (SFB 665), and by grants from the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of Japan to Y.M. and T.K. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z and Medical Research Council Hub grant G0900747 91070) for generating the sequencing data, John Broxholme for assisting with downloading BAM files from the 1000 Genomes Project, Kevin Leyden for performing flow cytometry, Christian Babbs for sharing DNA from Arabic

controls, and Kana Miyanagi for assisting in functional analysis of PGAP3. We would also like to thank all three families for their participation in this study and the 500 Whole-Genome Sequences Consortium and Illumina for use of the in-house database of variant calls for 274 individuals.

Received: September 25, 2013
Accepted: December 11, 2013
Published: January 16, 2014

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://www.1000genomes.org/
ANNOVAR, http://www.openbioinformatics.org/annovar/
Burrows-Wheeler Aligner, http://bio-bwa.sourceforge.net/
ClustalW, http://www.clustal.org/clustal2/
Coriell Cell Repositories, http://ccr.coriell.org/
dbSNP, http://www.ncbi.nlm.nih.gov/snp/
GATK, http://www.broadinstitute.org/gatk/index.php
GeneTalk, http://www.gene-talk.de
Ingenuity, https://variants.ingenuity.com/MutationsinPGAP3causeARID
HomozygosityMapper, http://www.homozygositymapper.org/
KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/
LifeScope, http://www.lifetechnologies.com/lifescope.html
MutationTaster, http://www.mutationtaster.org
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
Oxford Brain Abnormality Research Group, http://www.brainabnormalities.org.uk
Platypus, http://www.well.ox.ac.uk/platypus
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
SIFT, http://sift.jcvi.org/
UCSC Genome Browser, http://www.genome.ucsc.edu

References

1. Fujita, M., and Kinoshita, T. (2012). GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. Biochim. Biophys. Acta *1821*, 1050–1058.
2. McKean, D.M., and Niswander, L. (2012). Defects in GPI biosynthesis perturb Cripto signaling during forebrain development in two new mouse models of holoprosencephaly. Biol. Open *1*, 874–883.
3. Park, S., Lee, C., Sabharwal, P., Zhang, M., Meyers, C.L., and Sockanathan, S. (2013). GDE2 promotes neurogenesis by glycosylphosphatidylinositol-anchor cleavage of RECK. Science *339*, 324–328.
4. Jaeken, J. (2011). Congenital disorders of glycosylation (CDG): it's (nearly) all in it!. J. Inherit. Metab. Dis. *34*, 853–858.
5. Nozaki, M., Ohishi, K., Yamada, N., Kinoshita, T., Nagy, A., and Takeda, J. (1999). Developmental abnormalities of glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/loxP system. Lab. Invest. *79*, 293–299.
6. Johnston, J.J., Gropman, A.L., Sapp, J.C., Teer, J.K., Martin, J.M., Liu, C.F., Yuan, X., Ye, Z., Cheng, L., Brodsky, R.A., and Biesecker, L.G. (2012). The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. Am. J. Hum. Genet. *90*, 295–300.
7. Ng, B.G., Hackmann, K., Jones, M.A., Eroshkin, A.M., He, P., Wiliams, R., Bhide, S., Cantagrel, V., Gleeson, J.G., Paller, A.S., et al. (2012). Mutations in the glycosylphosphatidylinositol gene PIGL cause CHIME syndrome. Am. J. Hum. Genet. *90*, 685–688.
8. Almeida, A.M., Murakami, Y., Layton, D.M., Hillmen, P., Sellick, G.S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L., et al. (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat. Med. *12*, 846–851.
9. Maydan, G., Noyman, I., Har-Zahav, A., Neriah, Z.B., Pasmanik-Chor, M., Yeheskel, A., Albin-Kaplanski, A., Maya, I., Magal, N., Birk, E., et al. (2011). Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. J. Med. Genet. *48*, 383–389.
10. Kvarnung, M., Nilsson, D., Lindstrand, A., Korenke, G.C., Chiang, S.C., Blennow, E., Bergmann, M., Stödberg, T., Mäkitie, O., Anderlid, B.M., et al. (2013). A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. J. Med. Genet. *50*, 521–528.
11. Krawitz, P.M., Murakami, Y., Hecht, J., Krüger, U., Holder, S.E., Mortier, G.R., Delle Chiaie, B., De Baere, E., Thompson, M.D., Roscioli, T., et al. (2012). Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am. J. Hum. Genet. *91*, 146–151.
12. Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat. Genet. *42*, 827–829.
13. Murakami, Y., Kanzawa, N., Saito, K., Krawitz, P.M., Mundlos, S., Robinson, P.N., Karadimitris, A., Maeda, Y., and Kinoshita, T. (2012). Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome. J. Biol. Chem. *287*, 6318–6325.
14. Hansen, L., Tawamie, H., Murakami, Y., Mang, Y., ur Rehman, S., Buchert, R., Schaffer, S., Muhammad, S., Bak, M., Nöthen, M.M., et al. (2013). Hypomorphic mutations in PGAP2, encoding a GPI-anchor-remodeling protein, cause autosomal-recessive intellectual disability. Am. J. Hum. Genet. *92*, 575–583.
15. Krawitz, P.M., Murakami, Y., Rieß, A., Hietala, M., Krüger, U., Zhu, N., Kinoshita, T., Mundlos, S., Hecht, J., Robinson, P.N., and Horn, D. (2013). PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. Am. J. Hum. Genet. *92*, 584–589.
16. Tashima, Y., Taguchi, R., Murata, C., Ashida, H., Kinoshita, T., and Maeda, Y. (2006). PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Mol. Biol. Cell *17*, 1410–1420.
17. Maeda, Y., Tashima, Y., Houjou, T., Fujita, M., Yoko-o, T., Jigami, Y., Taguchi, R., and Kinoshita, T. (2007). Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol. Biol. Cell *18*, 1497–1506.
18. Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P.,

et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63.
19. Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.
20. Kamphans, T., and Krawitz, P.M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics *28*, 2515–2516.
21. Thompson, M.D., Roscioli, T., Marcelis, C., Nezarati, M.M., Stolte-Dijkstra, I., Sharom, F.J., Lu, P., Phillips, J.A., Sweeney, E., Robinson, P.N., et al. (2012). Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). Am. J. Med. Genet. A. *158A*, 553–558.
22. Heinrich, V., Kamphans, T., Stange, J., Parkhomchuk, D., Hecht, J., Dickhaus, T., Robinson, P.N., and Krawitz, P.M. (2013). Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. Genome Med. *5*, 69.
23. Kamphans, T., Sabri, P., Zhu, N., Heinrich, V., Mundlos, S., Robinson, P.N., Parkhomchuk, D., and Krawitz, P.M. (2013). Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS ONE *8*, e70151.
24. Murakami, H., Wang, Y., Hasuwa, H., Maeda, Y., Kinoshita, T., and Murakami, Y. (2012). Enhanced response of T lymphocytes from Pgap3 knockout mouse: Insight into roles of fatty acid remodeling of GPI anchored proteins. Biochem. Biophys. Res. Commun. *417*, 1235–1241.
25. Wang, Y., Murakami, Y., Yasui, T., Wakana, S., Kikutani, H., Kinoshita, T., and Maeda, Y. (2013). Significance of glycosylphosphatidylinositol-anchored protein enrichment in lipid rafts for the control of autoimmunity. J. Biol. Chem. *288*, 25490–25499.
26. Rudd, P.M., Morgan, B.P., Wormald, M.R., Harvey, D.J., van den Berg, C.W., Davis, S.J., Ferguson, M.A., and Dwek, R.A. (1998). The glycosylation of the complement regulatory protein, human erythrocyte CD59. Adv. Exp. Med. Biol. *435*, 153–162.