New Genes for Old Diseases: Using whole genome sequence to identify genetic causation in two families with unexplained Mendelian disease

Ben Faircloth

Supervised by Dr. David Markie & Professor Stephen Robertson

Acknowledgements

I would first like to thank my supervisors, Dr. David Markie of the Pathology department and Professor Stephen Robertson of the department of Women's and Children's health at the University of Otago. They were always available for advice and guidance when I needed them, and I couldn't have finished the project without their support. Your teaching and expertise while I was researching, and your many constructive comments throughout the writing process were invaluable.

I would also like to thank those in the clinical genetics group for their encouragement and suggestions, as well as the other postgraduate students in the department for their friendly support.

I want to thank my friends, particularly Will and Rebecca, for stopping me from becoming a hermit while I was writing and for their support and encouragement.

I also want to thank Shayma, my partner, for everything. You are my rock and you keep me sane.

Finally, I want to express my gratitude to family for their unfailing support and encouragement throughout the process of researching and writing this thesis. I couldn't have done it without you.

Abstract

Humans are afflicted by an enormous number of diseases with a genetic component, of which roughly 7,000 are thought to follow Mendelian inheritance. As these Mendelian diseases often have a large impact on normal development and quality of life, many studies are conducted using the affected individual's genetic sequence in an effort to determine what has changed in which gene to cause them. If successful, this can then aid in understanding the disease and how best to manage it, as well as furthering the effort towards understanding the function of every gene in the human genome. Many techniques are used to locate the causative changes, or variants, that are causing a given Mendelian disease, such as using exome sequencing to search the coding regions of the genome, or comparative genomic hybridization to arrays (array-CGH) to identify large deletions or duplications.

This analysis attempted to identify the genetic cause of unexplained disorders in two families, one affected with Otopalatodigital Syndrome Type 1 (OPD1) and the other affected with Larsen Syndrome. Both disorders are usually caused by variants in the genes that code for filamin proteins (*FLNA* and *FLNB* respectively). These families are atypical in that no causative variants in these genes had been found despite significant previous attempts. In the OPD1 family, the exons of the *FLNA* gene have previously been sequenced using the Sanger methodology and array-CGH had been performed, but no causal variant had been located. In the family affected by Larsen Syndrome, the Sanger methodology has been used to sequence across the exons and exon/intron boundaries of the *FLNB* gene and Multiplex Ligation-dependent Probe Amplification

(MLPA) has been performed over the *FLNB* gene. Ultimately, the whole exome of this family trio was examined, but no causal variant had been identified.

In this study, an analysis of whole genome sequence data was undertaken in an attempt to resolve the causation of the disorders in these two families. In the case of the family thought to be affected by OPD1, the child in fact had Rubinstein-Taybi syndrome; a disease with a similar phenotype but that is caused by mutations in the *CREBBP* gene, in this case a loss of a splice donor sequence at the beginning of exon 20. For the family affected by Larsen Syndrome multiple variants remain that could be causal, although none are particularly compelling. For this reason, it is not possible to definitively determine the cause of the disease and so no inferences about the genes with which *FLNB* interacts can be made. It is suggested that the analysis be repeated with further families as they become available, as the strength of the candidate genes would be greatly increased if they were also found in another family. Currently the cause of the disease in this family remains an enigma, despite the use of whole genome sequencing.

Table of Contents

List of Figures

List of Tables

List of abbreviations

Array-CGH	Array comparative genomic hybridization
ASXL3	Additional sex combs-like 3
CREBBP	CREB Binding protein
DNA	Deoxyribonucleic acid
ENCODE	Encyclopaedia of DNA elements
ExAC	Exome aggregation consortium
FLNA	Filamin A
FLNB	Filamin B
GATK	Genome analysis toolkit
GNAL	G protein subunit alpha L
HGVS	Human Genome Variation Society
IGV	Integrated genomics viewer
MLPA	Multiplex Ligation-dependent Probe Amplification
OPD1	Otopalatodigital syndrome 1
OPD2	Otopalatodigital syndrome 2
ORF	Open reading frame
PCR	Polymerase chain reaction
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
VQSR	Variant Quality Score Recalibration

1 Introduction

1.1 Finding Human Disease Genes: A Rationale

1.1.1 Genetics and disease

Humans are affected by enormous numbers of diseases caused by many different things, such as pathogenic bacteria, lack of an essential nutrient from the environment, and errors in cellular control and function due to changes in an individual's genetic code. While in some cases the cause of a disease is easily identified, in many cases it is extremely difficult. In particular, it can be incredibly hard to identify the gene or genes responsible for causing a genetic disease.

Identifying the gene or genes that cause a disease can be difficult because while it is known that there is at least one genetic change, or causal variant, that has occurred to cause the disease there are also many changes in the genetic code of the average individual that don't cause a disease, and the function of many genes is currently unknown. This makes identifying the cause of such genetic diseases hard because when there are changes, or variants, across many genes whose function is not well understood, it is impossible to say which of them is causing the disease. In addition to this, often a genetic disease is caused by variants in more than one gene which further compounds the issue, because to fully understand what is causing the disease it is necessary to understand the function of every gene involved and how they interact with one another. If it is possible to identify the exact genetic cause of a disease, there is the possibility of developing therapeutic interventions based on this knowledge, which if successful can lead to greatly improved quality of life for those affected.

1.1.1.1 Identifying candidate variants

Because of these issues, many techniques have been developed to help to identify the variations in an individual's genetic code that could be causing disease. Many tools, such as those in the Genome Analysis Toolkit (GATK), have been developed that can search an individual's available DNA sequence and compare it to a standard reference sequence to find any changes and attempt to assess whether they are likely to affect the normal function of a gene or genes. These changes are known as candidate variants and are defined as any change within the genetic code from single nucleotide polymorphisms (SNPs) to variations in copy number to large chromosomal rearrangements, that could be causing the affected individual's disease (Chong et al., 2015). However, when a disease is caused by variants in more than one gene it is often not possible to identify them, so most often molecular diagnosis is done when the disease follows Mendelian inheritance patterns, i.e. is caused by changes in a single gene.

1.1.1.2 Identifying gene function

Once candidate variants have been identified along with the gene whose function they are predicted to affect, it is necessary to determine the function of that gene and how a change in its function could be causing the disease phenotype seen. In some cases, such as if the variant is in a gene known to cause a disease that presents with a phenotype like that of the individual in question, this is trivial. However, the majority of the time the candidate variant is in a gene whose function is completely or partially unknown, or the disease the gene is known to cause presents a phenotype that could not be confused with the phenotype seen in the patient. This necessitates further research to elucidate the relevant function of the gene in question. This can be done in many ways, including utilising cellular assays or model organisms, and if successful

can confirm whether a candidate variant is causing the disease or not. However, when a disease is caused by changes in more than one gene, it is generally not possible to identify them all and understand how they contribute to the disease.

1.1.2 Mendelian disease vs complex disease

There are roughly 7,000 genetic diseases that follow Mendelian inheritance (i.e. are caused by changes in a single gene), although it is likely that there are more that are not yet defined (Boycott, Vanstone, Bulman, & MacKenzie, 2013). However, many diseases are due to changes in multiple genes in addition to environmental factors, such as cardiovascular disease (Feinleib et al., 1977). Diseases with a genetic basis that result from contributions from more than one gene, often with significant environmental contributions are called complex diseases. These can be difficult to study as the causative genes are often unknown, which makes it difficult to predict the phenotypic consequences of variants within them (Manolio et al., 2009). In addition to this, if multiple genes have a cumulative effect it is almost impossible to determine to what magnitude each gene is responsible for the disease (Alkuraya, 2016).

In contrast, the causes of diseases that follow Mendelian inheritance are relatively easy to determine, as there is usually one variant in one gene causing the disease. Such a variant can sometimes be clearly identified as the cause of the disease, and finding it can provide a great deal of value both clinically and scientifically.

1.1.3 The value of determining causal genes in Mendelian disease

1.1.3.1 Scientific value

Often the gene that is disrupted in Mendelian diseases is of unknown function. As one of the main goals of genetics based biomedicine is to determine the function of every gene in the human genome (Saleheen et al., 2017), determining which gene has been disrupted by a variant to cause a disorder is valuable. This is because after finding a causative variant in a gene it is possible to further elucidate that gene's function through the phenotype produced. For example, if the phenotype associated with a genetic disorder is that of decreased height, it can be surmised that the affected gene plays a role in the biological pathways that determine how tall someone is, either through playing a role in the production or regulation of growth hormone or through other means.

Similarly, although the phenotypic effects of a genetic disease are obvious in many cases, the exact biological cause of these effects can remain unknown for a long time due to the lack of understanding of the pathways involved. Using genetic techniques to study such diseases can lead to increased understanding of these pathways, as once the gene is known its protein product can be determined. Following this, further study can be undertaken to work out what function that protein performs and with what it interacts. This in turn can lead to possible ways to manage or treat the disease, improving the quality of life of anyone suffering from it.

1.1.3.2 Clinical value

In addition to enhancing our understanding of the biology of inherited diseases and their associated pathways, using genetic techniques to determine exactly what variants are causing the disease can help diagnose the disorder (Bell, 2004). Once a causative variant for a disease has been found, such variants can be looked for in cases where the physician is concerned that an individual is susceptible to genetic disease, or when there is a detrimental phenotype but its cause is unknown. This is particularly useful in the case of older fathers, as paternal age plays a role in the likelihood of genetic disorders occurring in their offspring (Veltman & Brunner, 2012). In such cases, a genetic screen that looks for any variants in genes known to cause Mendelian disorders can be conducted, which can inform the prospective parents of what to expect and provide options for minimising the effect of the disease.

Such diagnostic abilities are useful even when there is no adequate treatment, as is often the case with genetic diseases. This is because knowing what to expect and having management strategies available can reduce the burden of the disease for the affected individual and their family. In addition to this, having an accurate diagnosis often enables a more accurate prognosis, which can enable the best strategy of care to minimise harm.

As well as aiding in the diagnosis and informing the prognosis of individuals with a Mendelian disease, determining the causal variant can help inform the affected individuals' future reproductive choices, as well as enabling prenatal diagnosis if they do choose to have children.

Because of these positive outcomes from finding causal variants in affected individuals, a large amount of research is conducted with the aim of identifying them. There are many methods used to discover such genetic variants that cause disease, several of which are outlined below.

1.2 Current methods of obtaining sequence to find causal variants

1.2.1 Sequencing using the Sanger methodology

One of the best-known and most widely used methods for obtaining genetic sequences is the Sanger methodology. Used as the standard method of obtaining sequence for nearly 40 years, sequencing using the Sanger methodology gives reliable and high-quality results down to a single base pair, and is therefore very useful for proving the existence of variants that could be causing disease. When sequencing according to the Sanger methodology, a reaction is performed to construct a DNA molecule where each base is known. Each reaction contains DNA polymerase and free oligonucleotides of all four standard bases, in addition to smaller quantities of the bases labelled in such a way that the incorporation of one of them halts extension of the molecule (SenGupta & Cookson, 2010). This means that as DNA polymerase constructs the strand of DNA it will stop at random points whenever it uses one of the labelled bases. Each labelled base fluoresces a distinct colour, so when separated by size and viewed, the sequence of the DNA strand is readable. This method is robust and reliable, but is limited to fragments of ~900bp length due to deteriorating quality of the fluorescent traces (SenGupta & Cookson, 2010).

Although sequencing using the Sanger methodology gives very high-quality sequence data its usefulness for determining causal variants is limited. This is because it is very specific, as you need to design unique primers on either side of the section of DNA to be sequenced. Because of this specificity it is necessary to have candidate variants in mind before you design the sequencing polymerase chain reaction (PCR). It is also difficult and expensive to use the Sanger methodology to sequence all the coding regions of a candidate gene, as most genes are too long to be sequenced in a single reaction. This makes it necessary to sequence it in sections then search all the regions together, which can introduce errors (SenGupta & Cookson, 2010).

The Sanger methodology is still widely used to prove that variants found in exome and whole genome data are not false positives, for example artefacts of the alignment process. However, it is not often used to look at long sequences of DNA or multiple genes. It is necessary to use another technique to find candidate variants, then confirm them by sequencing using the Sanger methodology.

1.2.2 High throughput sequencing approaches

Over the past decade innovative technologies have made the use of high throughput sequencing by the average researcher a possibility, as higher quality results have become available in tandem with decreased costs. This has made it much more plausible to obtain large quantities of sequence data to be used to determine candidate variants, including the entirety of the coding regions of an individual's genome (the exome) and even the entire genome itself. This is particularly useful when a researcher is unsure of which gene is causing the phenotype they are seeing and so can't use the other methods listed. There are many companies with

competing techniques for producing this sequence, so here I will focus on the technology used by Illumina, as it is the most widely used and has been used in this study.

1.2.2.1 Illumina's method

First, the DNA to be investigated is fragmented into short pieces of a predetermined length (between 300 and 500 bases). The DNA used in this method can be up to the entire genomic sequence of an individual, but most often is the exome of an individual.

To use this method to sequence an exome it is first necessary to "capture" (selectively sequence) the parts of the genome included. To accomplish this, the fragments of genomic DNA are hybridized in solution to a set of labelled synthetic oligonucleotides that covers all the regions of interest. Following this, the fragments of interest can be separated using the labels on the synthetic oligonucleotides.

After fragmentation, adaptors are ligated onto both ends, along with a unique index sequence. These fragments are then loaded onto a flow cell, the surface of which has oligonucleotides complementary to the adaptors on the fragments. The fragments bind to the oligonucleotides, which are then amplified by bridge PCR to produce clusters that will provide sufficient signal for the rest of the analysis. Each flow cell has multiple lanes, each of which can produce millions of reads. This allows for the production of huge amounts of DNA in a short period of time.

Following cluster formation, a primer is attached to the free end of the bound fragments, after which fluorescent nucleotides are added with polymerase. The nucleotides are blocked in such a way that only one can be incorporated at a time. A laser then excites the fluorescent molecule and an image is obtained enabling determination of the base incorporated, similar to the

technique used in the Sanger methodology. After each cycle, the molecule blocking the addition of more nucleotides is removed so that another can be incorporated and the process repeats until the desired read length is reached. In the case of paired end reads, once the forward sequence has been determined the same is done for the reverse sequence. This process is called sequence-by-synthesis and is pictured in figure 1-1.

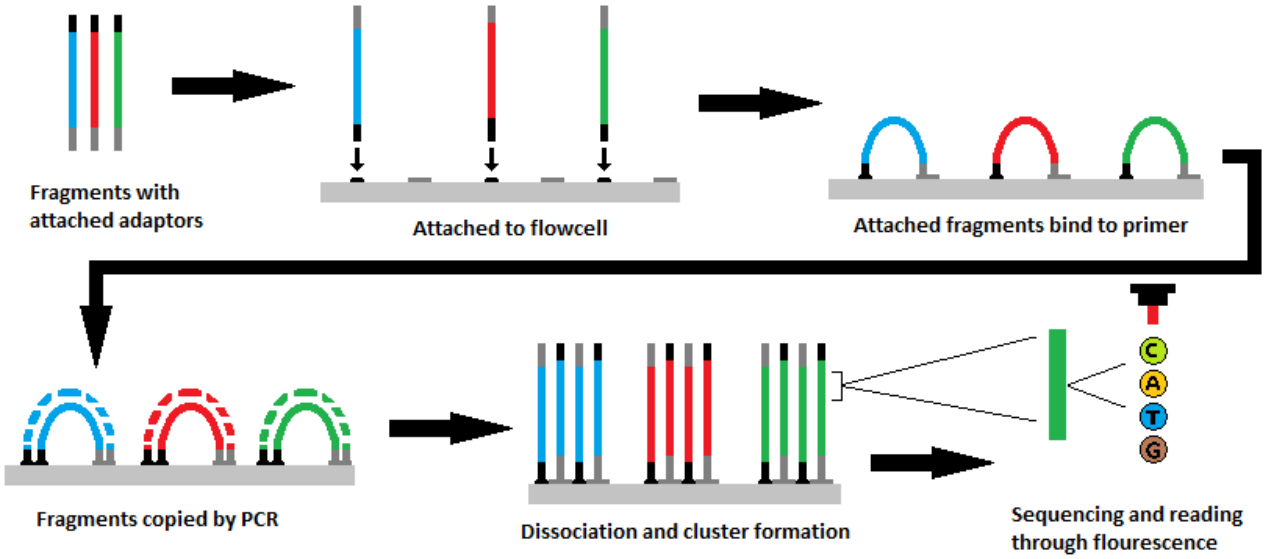

Figure 1-1: A depiction of Illumina's method for high throughput sequencing using sequence-by-synthesis. Adapted from http://www.3402bioinformaticsgroup.com/wp-content/uploads/2016/07/NGS.png, accessed 25/05/2017

1.2.2.2 Further processing

After the fragments are sequenced they are separated in silico using the unique index sequences. The fragments are then aligned back to a reference sequence, producing DNA sequence for as much of the initially fragmented DNA as possible. Paired end reads can be used to resolve areas where the alignment is ambiguous, as they are of a known length. The

differences between this sequence and the reference sequence can then be used to identify candidate variants that could be causing the disease.

After alignment, duplicated reads are removed as they can lead to incorrect conclusions about the biological signal of variants (Dozmorov et al., 2015). For example, if an error occurs in the initial extension of a fragment causing an apparent variant, and that fragment is duplicated many times, it can lead to a false conclusion that the individual in question has a variant at that position. Additionally, the confidence with which each base was identified during the sequencing stage is used to give them a quality score. The likelihood of these variants being real, or their quality, can then be measured using factors including the base quality scores. This list of variants can then be annotated with useful information, such as the frequency at which they are found in a control population (for example the Exome Aggregation Consortium, ExAC, which is a collection of the exomes of over 60,000 unrelated individuals) and their predicted effect on the protein produced by the gene they are in. These steps are usually performed using a workflow which contains tools to accomplish them, such as the Genome Analysis Toolkit (GATK). This approach is very useful for finding short variants, such as single nucleotide variants (SNVs) and small indels (the insertion or deletion of a small number of bases), most often in the exome of an individual.

1.2.2.3 Filtering the output of high throughput sequencing

After annotation, the output of the workflow can be filtered using the results of the annotation to find variants that could be causal within the extensive list of variants produced. For example, you can remove all variants that are present at a population allele frequency higher than a cut-

off point. This is useful because if a variant is present at a relatively common frequency in a population of unaffected individuals then it is unlikely to be causing a rare and highly penetrant disease phenotype. Other useful ways of filtering are based on the predicted effect of a variant, as some types of variant are more likely to affect function than others, such as frameshift variants or a variant that destroys a splice site. Using these filters is necessary to reduce the number of candidate variants to a manageable number, as the number of variants in the average individual's exome is huge, and there are more in the whole genome.

1.2.2.4 Trio design

When filtering to find causal variants it is useful to use trio design. Trio design is using the sequence of the affected individual's parents in addition to their own, as this enables the construction of inheritance models which means that particular types of variant can be filtered for. For example, using trio design you can construct a model that locates and retains variants present at positions where the affected child has a variant when compared to the reference sequence, but the parents do not, and removes everything else, i.e. *de novo* variants. Other models that can be used include recessive models for homozygous and compound heterozygous traits.

1.2.2.5 Exome data

The exome is the subset of the genome contained within exons, i.e. the coding part of the genome, and most of the time will contain the disease-causing variant (Choi et al., 2009). Sequencing the whole exome of an individual has become routine, as evidenced by the substantial number of exomes in the ExAC (Exome Aggregation Consortium) database, which

contains the exomes of 60,706 individuals (URL http://exac.broadinstitute.org/). An exome sequence gives good coverage at very high resolution, down to individual bases. Utilising the filtering techniques outlined above with exome data gives a very good chance of finding the causative variant, but only if it is in a coding region of the genome.

One disadvantage of using exome data is that there is no straightforward way to find structural variants in the data, as there is no simple way to filter for them. In addition to this, exome data has many false positives due to errors in annotation and alignment, so it is necessary to check the existence of any possible causative variants by sequencing using the Sanger methodology, which can be expensive if there are multiple candidate variants that remain after filtering. Another issue when using exome data is that the regions of the genome included vary depending on which capture platform you use, as different companies can include slightly different regions of the genome in their kits. As such it is possible that some variants could be missed during analysis.

1.2.2.6 Whole genome data

Sometimes a researcher may choose to sequence the whole genome directly, although the cost is higher. Alternatively, if after examining exome data no variant likely to be causing the disease in the individual under investigation has been found, the entire genome may then be sequenced to obtain any candidate variants that may have been missed previously. Sequencing entire genomes using high throughput methods is expensive and analysing the data produced requires a lot of computing power, but it will include all coding and non-coding regions of that

individual's DNA down to base-pair resolution. This means that if there is a causative variant in the individual, it will likely be in the dataset.

One problem with whole genome sequencing is that it is still possible to miss the causative variant even with the huge amount of data obtained. For example, it is often difficult to assess the significance of variants in non-coding regions because of the current general lack of knowledge about regulatory regions. In addition to this, whole genome data contains many false positive results because of misalignment or incorrect annotation (Steward et al., 2017). Because of this, it is necessary to validate the variants by sequencing using the Sanger methodology.

1.2.1 Methods for finding structural variants

1.2.1.1 Array-CGH

One of the main ways structural variants, especially deletions and duplications, can be found is by comparative genomic hybridization to arrays (CGH) (Carter, 2007). Using CGH arrays it is possible to perform a genome-wide analysis of copy number in one experiment. An array of overlapping target sequences, most often composed of known SNPs that are accurately mapped to the human genome, is developed through creating synthetic oligonucleotides, which are then attached to beads. These beads are then arrayed in wells across a slide. Following this, sample genomic DNA is added and allowed to hybridize to the array of sequences on the slide. The number of sample molecules hybridized is then counted and structural variants detected based on the number of hybridized molecules observed versus the number expected. This is relatively quick and easy to perform after the oligonucleotides have been created, but there are issues

with long repeat sequences and cross-hybridization (Ylstra, van den Ijssel, Carvalho, Brakenhoff, & Meijer, 2006). In addition, the resolution of this technique is dependent on the map density of the markers used, which means that in order to ensure that no variants are missed by the analysis it is necessary to use huge numbers of markers.

1.2.1.2 MLPA

Another way structural variants can be discovered in the human genome is by using MLPA (Multiplex Ligation-dependent Probe Amplification). MLPA is a method used to detect abnormal copy numbers across a small number (up to 40) of genomic DNA or RNA sequences (Schouten et al., 2002). In MLPA, probes added to the samples are amplified, rather than the sample nucleic acids themselves. Each probe consists of two oligonucleotides that hybridize to a target sequence next to each other, after which they can be ligated to each other. Ligated probes have identical sequences at their 5' and 3' ends, which means that all probes from a test can be amplified simultaneously in a PCR with only one pair of primers.

MLPA is very sensitive, and can differentiate between sequences that differ in the deletion or duplication of only one copy of one exon, unlike most other techniques used to find variance in copy number. For example, Southern blots would find most deletions and duplications, but they take a lot of time and only find changes within the probes used, as well as requiring a large quantity of DNA (Schouten et al., 2002). In addition, heterozygous deletions and duplications in human DNA aren't found through methods based on PCR amplification of genomic DNA, as a normal allele is also present in such samples. Other advantages of using MLPA are that it uses a small quantity of DNA (~20ng), is cheaper than other methods and it takes less time to perform

(Schouten et al., 2002). The equipment necessary to perform MLPA is also widespread and present in many laboratories as standard.

However, this technique is limited to using 40 target sequences per reaction, so it cannot be used unless you already have candidates with a high index of suspicion, similar to when using the Sanger methodology. Also, in order to perform MLPA you need to create the probes in the first place, which is time consuming, although once this is done generally enough probes are produced for a large number of reactions (Schouten et al., 2002).

1.2.1.3 Genomic data

As genomic data contains sequence information for the entire genome, usually mapped using paired end reads, it can be used to find structural variants in addition to SNVs and indels. This is because if one of a pair of reads maps to a different location than the other half of the pair they will be flagged as discordant in the alignment data. Many discordant pairs of reads at the same position on the genome is suggestive of a structural variant. Another indication of the presence of a structural variant in genomic data is the average number of reads across an area. If the depth of coverage is significantly lower or higher than the average across the rest of the sequence data, it is possible that there is a structural variant at that location. Finally, during alignment a single read can be split to map to non-adjacent locations, which suggests that the read crosses the breakpoint of a structural variant in that individual.

However, although it is possible to use these indicators to suggest a structural variant using genome data it is difficult to develop a way to filter for them with high confidence. This is because lots of noise is generated by areas with many repetitive sequences, where reads will

map as discordant or split during alignment by mistake. In addition, there is a great deal of variability in the depth of coverage across the genome which can give a misleading impression that a deletion or insertion is causing decreased or increased coverage, when in fact it is due to chance. Because of these false positives, stringent filtering methods are required to use genomic data to search for structural variants as a consequence of which both the specificity and sensitivity of detecting true variants is reduced compared to the detection of SNVs and short indels.

One method that can be used to identify structural variants from genomic sequence data is the GenomeStrIP SV workflow (Handsaker, Korn, Nemesh, & McCarroll, 2011). This workflow can only identify deletions, and it does so by using discordance between pairs of reads to suggest potential deletion breakpoints, after which it utilises depth of coverage to assess locations where a deletion may have occurred. Another workflow that attempts to characterise structural variants, in this case of all kinds, is Lumpy (Layer, Chiang, Quinlan, & Hall, 2014). Lumpy uses split reads and discordant reads in addition to the orientation of the paired reads to attempt to discover all structural variants. These methods of filtering the genomic data ultimately provide some candidate structural variants, although with a reduced sensitivity and specificity relative to when detecting short indels and SNVs.

1.3 Families studied in this thesis and their conditions

1.3.1 The families

In this thesis, genomic sequences are utilised to attempt to determine the genetic cause of unexplained disorders in an affected child from two families, one affected with Otopalatodigital Syndrome Type 1 (OPD1) and the other affected with Larsen Syndrome. Both are disorders usually caused by variants in the genes that code for filamin proteins (*FLNA* and *FLNB* respectively). In this investigation, genomic data was used because in these families no causative variants have been found despite a thorough investigation. In the OPD1 family, all of the exons of the *FLNA* gene have previously been sequenced using the Sanger methodology and array-CGH has been performed. The Larsen syndrome family has previously been assessed using the Sanger methodology to sequence across the exons and exon/intron boundaries of the *FLNB* gene, as well as MLPA over the *FLNB* gene. Ultimately, the whole exome of this family trio was investigated, but still no causal variant was identified using standard filtering approaches.

1.3.2 The Filamins and their genes

1.3.2.1 The filamin proteins and their structure

The filamins are a family of actin binding proteins originally purified in 1974 by a group attempting to understand phagocytosis, who found that in their extracts there was actin, myosin and an unknown high molecular weight protein that bound to actin (Hartwig & Stossel, 1975). In mammals, the Filamin family has three members, filamin A, filamin B and filamin C (Feng & Walsh, 2004). The genetic organisation of the three genes that encode these proteins

(*FLNA*, *FLNB* and *FLNC*) are highly conserved, and all three are expressed at high levels during development (Feng & Walsh, 2004). The filamin proteins themselves show 60-80% homology across their sequence except for the hinge regions, which are more diverse (Feng & Walsh, 2004).

The structure of filamin A is characterised by a highly-conserved N- terminal actin binding domain which is followed by 24 β-pleated 'filamin repeats', within which are two hinge regions (between repeats 15-16 and 23-24) that give the protein some flexibility (Robertson, 2005). Filamin A forms dimers at the C-terminal of the protein, at the 24th β-pleated repeat (Gorlin et al., 1990). This dimer forming C-terminal is critical to the function of the protein as, in addition to linking the dimer, many proteins that interact with Filamin A do so at the C-terminal (Robertson, 2005). Variants in *FLNA* can cause severe phenotypes, up to and including lethality (Robertson, 2005). For example, variants in the actin-binding domain or repeats 3, 10 or 14/15 cause a spectrum of disorders including OPD-1 (Robertson et al., 2003). The structure of filamin A is shown in figure 1-2, and filamins B and C have similar structures.

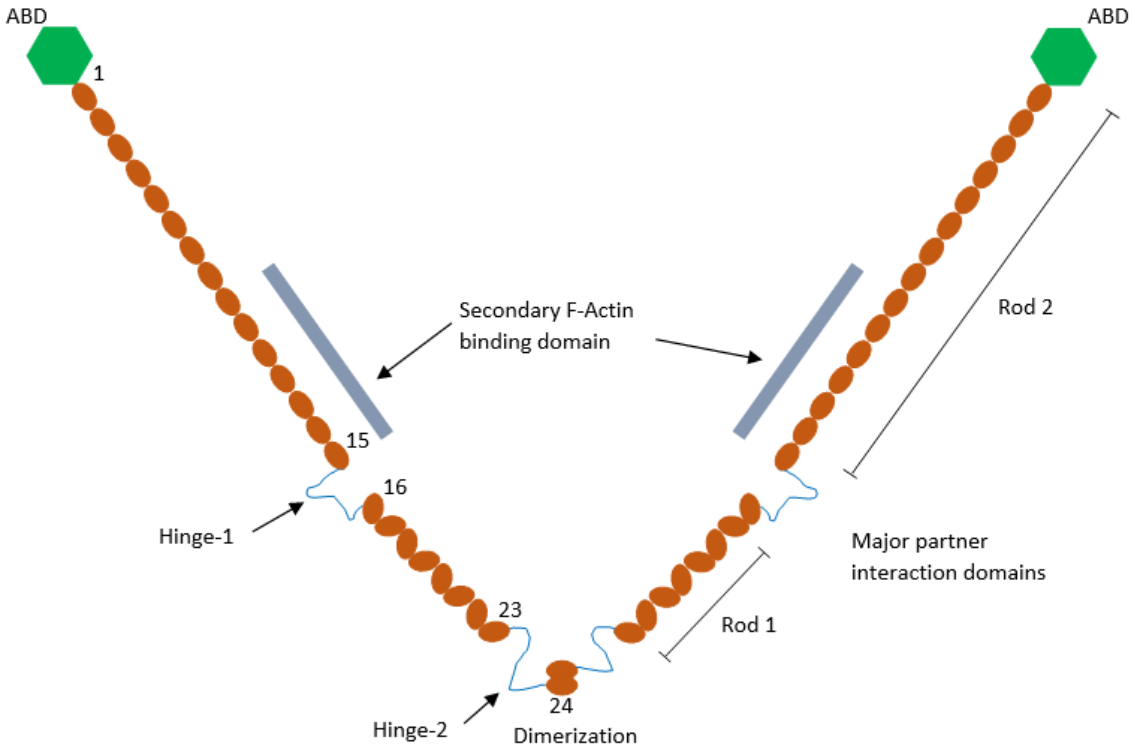

Figure 1-2: The structure of Filamin A. Shown is a dimer connected at the C-terminal at the 24th β-pleated repeat. The Actin Binding Domain (ABD) is at the N-terminal of each half of the dimer and the hinge regions are between repeats 15-16 and 23-24. (adapted from Stossel et al., 2001)

1.3.2.2 The function of the filamins

The filamin proteins operate by cross-linking actin filaments into networks or bundles to form a 3D structure, which then aid in cell motility or otherwise alter cellular structure (Stossel et al., 2001). In addition to binding actin, filamins have been shown to bind to upwards of 20 other macromolecules, including β-integrins and intracellular signalling components (Stossel et al., 2001). Although the physiological effects of some of these interactions have not yet been discovered, almost all of those known have a role in providing mechanical stability in the cell membrane as well as the maintenance of the connections between cells (Stossel et al., 2001).

Filamins are thought to facilitate the occurrence of cellular processes, particularly those that require the polymerization of actin, by bringing together these diverse macromolecules. Variants in the genes that encode these filamin proteins are associated with a wide range of diseases, including Melnick-Needles Syndrome, OPD1, OPD2 and Larsen syndrome (Robertson et al., 2003).

1.3.2.3 Otopalatodigital Syndrome Type 1 (OPD1)

OPD1 is an X-linked dominant genetic disease caused by gain of function variants in the *FLNA* gene, and is classified as a rare condition with an incidence of <1 in 100,000 individuals (Robertson, 2007). Its phenotype is the mildest of the OPD spectrum disorders and its characteristic features include dwarfism, spatulate fingers and toes, loss of hearing due to malformation of the ossicles in the ear, cleft palate and facial dysmorphisms such as a broad nasal root and frontal bossing (Dudding, 1967). There is no intellectual impairment in OPD1 although there can be in other OPD spectrum disorders (Robertson, 2007).

1.3.2.4 Larsen Syndrome

Larsen syndrome is an autosomal dominant osteochondrodysplasia normally caused by variants in Filamin B on chromosome 3 (Bicknell et al., 2007), but in rare cases it can be caused by recessive variants in the *B3GAT3* gene. Larsen syndrome is characterised by congenital large-joint dislocations and characteristic craniofacial abnormalities, including a prominent forehead and a flattened midface (Bicknell et al., 2007). In addition to these features, sufferers of Larsen syndrome also have spinal abnormalities such as scoliosis. Those affected by this syndrome often have hearing loss caused by malformed auditory ossicles (Bicknell et al., 2007).

1.4 The aims of this thesis

This thesis aimed to determine the causative variants of these diseases in each of the selected families. If successful, it is possible that the results will greatly enhance our understanding of these diseases and their causes. In addition, it is likely that since the phenotypes are similar to those of typical OPD1 and Larsen syndrome, finding causal variants in other genes will indicate that those genes or their products interact with filamin or the filamin genes in some way. This means that finding these variants could elucidate more of the pathway of those genes and with what their products interact, giving us a greater understanding of the associated cellular processes.

In addition, this thesis also aims to assess the use of whole genome sequence for identifying causal genes in cases that have otherwise been thoroughly investigated, as well as to implement and test the use of prediction of the presence of large deletions from whole genome sequence.

2 Methods

2.1 DNA Samples and processing

2.1.1 Obtaining samples

Information regarding patients and their families was obtained from Professor Stephen Robertson, who had collected cases from collaborating centres. DNA had previously been extracted from peripheral blood samples and coded with their individual ID numbers. These samples were available from archival stocks in the Robertson Laboratory. The ethics approval under which these samples were used was 13/STH/56, Study title: Genetic and functional studies into the causation of congenital malformations.

2.1.2 Whole Genome DNA Sequencing

DNA samples were sent to the Kinghorn Centre in Sydney for whole genome sequencing. Sequencing libraries were constructed with the Truseq Nano DNA Library Preparation kit and run on an Illumina X Ten system using pair end sequencing with around 150 cycles. Subsequent reads were available as paired, gz-compressed, fastq format files for each individual.

2.1.3 Sequence Processing for Alignment

The compressed fastq files for each individual were checked for quality parameters using FastQC v0.11.2 (URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then processed for alignment and variant calling according to the Genome Analysis Toolkit (GATK) Best Practice guidelines. The reads from each individual were aligned to the human reference

genome GRCh37 (URL https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37) using BWA v0.7.13 (URL https://github.com/lh3/bwa) with the mem algorithm. The primary alignment was then sorted and converted to bam format using the SortSam tool v2.3.0 from Picard (Li et al., 2009). Duplicated reads were identified with Picard MarkDuplicates (URL https://broadinstitute.github.io/picard/). Following this, tools from GATK 3.5 were used to realign around indels and perform Base Quality Score Recalibration to produce the completed alignment (bam format) for each individual.

2.1.4 Confirming Identity and Family Relationships

To confirm the identity of the individual under examination the GATK GenotypeConcordance tool was used according to the documentation online (URL https://software.broadinstitute.org /gatk/documentation/) to look for concordance between current (whole genome) and previous (exome) data, where exome data was available.

To achieve this, a small "fingerprint" variant file (vcf format) was generated using GATK HaplotypeCaller in genotype mode to genotype selected loci with common polymorphisms in most populations. The collection of loci used was the HapMap 3.3 markers, with overall allele frequencies in that population between 0.3 and 0.7. A threshold of 98% genotype concordance between exome and whole genome data was used to confirm identity.

Following confirmation of identity, it was necessary to confirm that the relationships between the individuals in question were as recorded. Therefore, the GATK SelectVariants tool (version 3.6-0-g89b7209, 2016) with the –mv option was used to produce a file containing Mendelian

violations from each trio. If the relationships described in the pedigree are correct a low number of violations is expected.

2.1.5 Variant calling and annotation

2.1.5.1 Determining gender

The gender of each sample was determined from the sequence data in the bam file by calculating the number of X and Y chromosomes using depth of coverage comparisons between the true X and Y chromosomes and an autosome (chromosome 1) with the GATK DepthOfCoverage tool. Depth of coverage estimates were restricted to coding regions (using the Agilent SureSelect All Exon version 5 manifest) to minimise the uncertainty produced by largely non-coding polymorphic repeat variation on the Y chromosome.

2.1.5.2 Variant calling

In the absence of conflict between the calculated and the reported gender, individual variant calling proceeded using the GATK HaplotypeCaller tool to produce a genomic vcf format file. The default ploidy of 2 was used for all regions except the true X region and the Y chromosome in males, where a ploidy of 1 was used to produce more representative genotype calls.

2.1.5.3 Compiling and quality control

Following this, multiple samples were incorporated into a multi-sample vcf file using GATK JointGenotyper and then processed through GATK LeftAlignandTrimVariants. The multi-sample vcf file was then annotated with a variety of internally calculated metrics using GATK AnnotateVariants, and GATK VariantRecalibrator was used to perform Variant Quality Score Recalibration (VQSR), which separates out variants into four groups, called truth sensitivity

tranches, by the probability that they are real. The highest tranche has the least false positives but is not very sensitive, and each subsequent tranche is more sensitive but has more false positives. The bands used in this study were 90.0 (estimated to identify 90% of true positives, missing 10% of them but not including almost all false positives), 99.0, 99.90 and 100 (i.e. all variants included). If a very high accuracy call set is required, then all variants outside the highest tranche can be removed, but if a more complete call set is a high priority then the lower tranches can be included.

2.1.5.4 Refinement

For the genotype refinement workflow, the autosomes were processed with GATK CalculateGenotypePosteriors to recalculate genotype quality after considering information from 1000 genomes and from the respective pedigrees. GATK VariantFiltration was also used to mark genotypes with a genotype quality less than 20, and GATK VariantAnnotator was used to mark possible *de novo* variants. The sex chromosomes were not processed through the genotype refinement workflow as the tools do not work with non-diploid regions, but the unprocessed variants from these regions were reintegrated into the resulting multisample vcf file using GATK CombineVariants.

2.1.5.5 Annotation

Gene context annotation and impact prediction was added using SnpEff 4.2 (URL http://snpeff.sourceforge.net/ SnpEff_manual.html) using Ensembl transcript data. SnpSift 4.2 (URL http://snpeff.source forge.net/SnpSift.html) was used to add phastCons (URL http://compgen.cshl.edu/phast /phastCons-HOWTO.html) data and selected information from

dbNSFP 2.9 (URL http://snpeff.sourceforge.net/SnpSift.html#dbNSFP). GATK VariantAnnotator was used to transfer population allele frequencies from ExAC v0.3 (Lek et al., 2016) and (Auton et al., 2015), and clinical annotations from ClinVar 2016_01_04 (https://www.ncbi.nlm.nih.gov/clinvar/intro/). These annotations were used selectively to filter variants for candidates.

2.1.5.6 Deletions

2.1.5.6.1 The GenomeStrIP SV workflow

As GATK HaplotypeCaller cannot be used to locate structural variants, it was necessary to use GenomeStrIP SV to find deletions (Handsaker et al., 2011). The GenomeStrIP SV release used was SV toolkit_2.00.1636, which has embedded within it GATK v3.5. GenomeStrIP SV requires at least 20 samples to be processed together, so the entire cohort of 145 whole genome samples available from other analyses in the clinical genetics group was processed together to produce a multisample vcf describing deletions in the cohort. These 145 individuals did not have the same disease as the individuals used in this analysis, nor were they relatives of patients with the same disorder. After the discovery and genotyping phases of the workflow default filters were used for the rest of the process. This workflow does not find all structural variants (only deletions). In addition, adequate annotation of allele frequencies cannot currently be obtained from sources such as 1000 genomes because the start and end positions of the deletions are approximate. This means that depending on the source you use, the deletions will be in slightly different positions and therefore not recognised as the same deletion.

2.1.5.6.2 Population construction

The ambiguity in the breakpoints of the deletions makes it necessary to use individuals processed together throughout the genotyping phase of the SV workflow as the control population, so that any deletions have the same definitions. To accomplish this, a file was created using the collection of sequences available from the clinical genetics group that contained the whole genome data from unrelated individuals that do not have the disease of interest. This file contained the information from 87 individuals, giving an effective population size of 87. The values for allele number, count and frequency it contains were then used to annotate the file containing the family of interest so that similar filters to those used for short variants could be applied to deletions.

2.1.1 Coverage and variant metrics

Depth of coverage statistics for each individual were obtained from the bam files using the GATK DepthOfCoverage tool and a description of the SNPs and Indels was obtained in a variant call file using the Picard CollectVariantCallingMetrics tool (see Appendix one for results).

2.2 Filtering Tools

When using whole genome data huge numbers of variants are discovered, particularly in intronic and intergenic regions. This makes it necessary to reduce the number of candidate variants by applying different parameters a variant must meet to remain a candidate for causing disease. There are a wide variety of tools available to perform such filtering that can be used independently, in combination, or not at all, depending on the inheritance model being investigated. Those used in this thesis are described in this section.

2.2.1 Population frequency filtering

2.2.1.1 Population frequency

If a variant is present at an allele frequency higher than could account for the disease prevalence then it is unlikely to be causing the disease, which makes it useful to filter out variants present at a frequency higher than a set threshold in the general population. The selected allele frequency threshold depends on the penetrance of the disease, the inheritance model being investigated and the number of distinct variants in the population that cause the disease. To accomplish this, SnpSift was used to filter on allele frequencies in the ExAC and 1000 genomes populations. In this case, the threshold frequency used was relatively conservative, as the more stringent the filter applied the more likely it is that the causative variant will be removed, so a conservative filter was applied then candidate variants were validated by sequencing using the Sanger methodology.

2.2.1.1.1 The ExAC database

One of the databases used to obtain the population frequency of the variants was the ExAC (Exome Aggregation Consortium) online database. This database contains the exome sequence of 60,706 unrelated individuals from multiple populations (exact numbers for population size are located at http://exac.broadinstitute.org/faq). The ExAC database contains a great deal of data on short variants, such as their location, gene and overall frequency in the combined populations. This makes it a very good resource for filtering out common variants, but as the disease in question may not be completely penetrant or an affected individual could have been included in the database inadvertently, it is necessary to filter for variants that are sufficiently

rare in this database as opposed to simply not present. Another reason such filtering is necessary is because the disease could be recessive, in which case heterozygotes may be present.

2.2.1.1.2 The 1000 genomes database

The other database used to calculate the population frequency of the variants was the 1000 genomes database. This database contains the genome sequence of 2504 individuals from multiple populations. The database contains information on short variants such as their overall frequency in the populations, location and gene. This database was used as it allowed filtering in the non-coding regions of the sequences used, which cannot be done with ExAC since it contains only exome data.

2.2.1.2 Separating individuals of interest from the cohort

When filtering whole genome data for candidate variants it is possible to use the vcf containing the whole cohort, but for convenience it is more common to separate out the individuals of interest. This was accomplished using GATK SelectVariants version 3.6-0-g89b7209 to extract variant data for selected groups of individuals into a separate vcf file, which was then used for further filtering.

2.2.1.3 Variant quality

2.2.1.3.1 Genotype quality

Variants are assigned a genotype quality score that reflects the confidence that they are correct, and a low quality score is an indication that the call may be equivocal on the basis of there being a low read depth or high allelic bias at that base. Variants that have low quality scores can be

filtered out if needed using the SelectVariants tool with the -select option, using the getGQ criterion. Often the genotype quality filter applied using Selectvariants is conservative, to reduce the risk of accidentally filtering out the causal variant.

2.2.1.3.2 Truth sensitivity

Another measure of quality that can be used to filter variants is the truth sensitivity tranche into which they fall. The truth sensitivity tranches separate out variants into four groups by the probability that they are real, and variants that fall into the lowest truth sensitivity tranche (VQSRTrancheSNP99.90to100.00) can be excluded using the -ef option in the GATK SelectVariants tool. This tranche has the highest rate of false positives, but when filtering out variants that fall into this tranche 0.1% of true positives will also be removed. In addition, this filter cannot be used for deletions as truth sensitivity scores are not calculated in that workflow.

2.2.1.4 Variant impact

SnpEff impact scores were used to filter out variants predicted to have little or no effect on the function of a gene. This was accomplished by using SnpSift to filter such that only variants given a moderate or high impact were kept. Variants given a high or moderate impact score are usually missense or truncating variants. Filtering for impact scores reduces the area of the genome included in the analysis to an area where a variant has an effect on translated products but in doing so increases the likelihood of the variants remaining after filtering having a phenotypic effect.

2.2.2 Inheritance model filtering

2.2.2.1 The *de novo* model

After separation of the trio of interest, a conservative list of *de novo* variants in the affected individual was created through filtering of the .vcf file using SnpSift. To filter for autosomal *de novo* variants, loci where the parents were both homozygous for the reference sequence, and the affected child showed a variant, were filtered for simultaneously. Using the SnpSift tool, this is done by keeping variants annotated as loConfDeNovo and hiConfDeNovo by the Genotype Refinement workflow in the affected child. As the workflow does not annotate *de novo* variants on the X chromosome it was necessary to filter for them separately. This was done using GATK SelectVariants to select bi-allelic variants where the non-reference allele was present in the affected individual, but not in either parent. GATK CatVariants was then used to combine the two resulting vcf files of *de novo* variants into one list of candidate *de novo* variants. It is possible that when using whole genome data too many *de novo* candidates to properly assess are found. In such cases, SnpSift was then used to filter for the predicted impact of the *de novo's* found. Following this, SnpSift was also used to filter for variants with a low enough allele frequency in both the ExAC and 1000 Genomes databases.

2.2.2.2 The recessive model

2.2.2.2.1 Homozygous

To create a list of possible homozygous variants, SelectVariants was used to simultaneously filter for locations where the affected individual was homozygous variant and both the mother and father were heterozygous. Low quality variants were excluded using GATK SelectVariants

with the –select option, using the getGQ criterion. SnpSift was then used to filter for variants of sufficient rarity in both the ExAC and 1000 Genomes databases. Following this, the remaining candidate variants were filtered on their predicted impact, again using SnpSift.

2.2.2.2.2 Compound Heterozygous

To find compound heterozygosity it was necessary to filter for genes where there were two candidate variants in the same gene, one inherited from the father and the other from the mother. This was accomplished by first using SelectVariants to simultaneously filter for locations where the child and father were heterozygous but the mother was homozygous reference. Then SelectVariants was used to filter for locations where the child and mother were heterozygous but the father was homozygous reference.

Following this, a list of the genes that contained the remaining variants from the father was created, as well as one containing the variants remaining from the mother. To do this the annotation of transcript names created earlier by SnpEff 4.2 was used. After creating these lists, they were compared and any gene that contained at least two candidate variants in the child, where one candidate was inherited from the father and the other from the mother, was placed into a final list of genes that could be affected by compound heterozygosity. The final list was then further filtered to only include rare variants using SnpSift, after which low quality variants were excluded using both the –ef option and the –select option with the getGQ criterion in the SelectVariants tool.

2.2.3 Deletions

After annotation through the GenomeStrIP SV workflow and construction of the population file, a file containing the data for just the trio of interest was created. This file contained information on deletions in the individuals in the same way that the files created through GATK HaplotypeCaller contained information on short variants. These files were then used to look for deletions that met the criteria for the same models as the short variants.

2.2.3.1 *de novo* deletions

To filter for *de novo* deletions, the vcf containing information on the deletions present in the trio was filtered as for *de novo* short variants. This involved filtering for locations where the child carried a deletion compared to the reference sequence but both parents did not, using GATK SelectVariants, followed by using SnpSift to filter out variants present at too high a frequency in the control population.

2.2.3.2 Recessive deletions

2.2.3.2.1 Population frequency filtering for deletions

As there is some ambiguity in the breakpoints of deletion variants it is necessary to construct a population of individuals processed together during genotyping for use as the control population. Therefore, a file was created that contained from individuals in the cohort of available sequences that are unaffected by the disease of interest. This file was then used to annotate the file containing the family of interest so that similar filters used for short variants could be applied to deletions. However, as the number of individuals in the control population file was small, the allele frequencies generated using it were crude and inaccurate. Because of

this, instead of using allele frequencies in the control population, allele counts were used, where if a variant was present more than once in the control population it would be excluded as a candidate for being too common. This filter was applied using the SnpSift tool.

2.2.3.2.2 Homozygous

In order to locate any homozygous deletions in the child, a filter that kept only locations where the child was homozygous and did not match the reference sequence was applied using GATK SelectVariants. This list was then further filtered to only include rare deletions using SnpSift. In addition to filtering for rarity, candidates that were not of sufficient quality were filtered out using the –ef option in the SelectVariants tool.

2.2.3.2.3 Compound heterozygous

To find locations where a deletion could be causing compound heterozygosity, it was necessary to filter for deletions that were close enough to a gene that it could affect its function, where that gene also had another candidate variant. This was accomplished by filtering for heterozygous variants that could affect gene function within one megabase (1,000,000bp) of either side of the approximate breakpoint of a heterozygous deletion. One megabase was chosen as the distance a deletion could be from a variant while still potentially affecting the gene containing the variant, because previous literature has suggested that enhancers can be up to a megabase from the gene they affect (Pennacchio *et. al.*, 2013). Any variants located within that distance were then further filtered to remove variants that were too common in the ExAC and 1000 Genomes databases using SnpSift. In addition to filtering for rarity, variants that were not of sufficient quality were filtered out using the –ef option in the SelectVariants tool.

2.3 Selection and vetting of candidates

2.3.1 The VCF file

Following filtering, a relatively short list of variants remained in a vcf from each model, including both real variants and some false positives. These files were manually examined to determine which of the remaining variants should be considered candidates for causality. Factors used to decide included the gene the variant was in, realistic values for allele number and a low frequency in the populations used. In addition, as it is known that the workflow has difficulties aligning around repeat elements such as microsatellites, variants in those areas were considered less likely to be real.

2.3.2 IGV browser

Another way the list of variants was vetted to remove variants not likely to be causative was using IGV (Integrated Genomics Viewer) to examine depth of coverage and read quality, as well as to see if the values in the vcf file for allele number matched up (Thorvaldsdottir et al., 2013). IGV shows if the pairs of reads are discordant, a histogram of the number of reads at any given base, and the base call for each position. Figure 2-1 is an example of a typical screen in IGV. If IGV showed low quality reads, low coverage, that one allele was much more common than the other, or that the variant was within (or at the end of) a repeat element it was considered less of a candidate for causality.

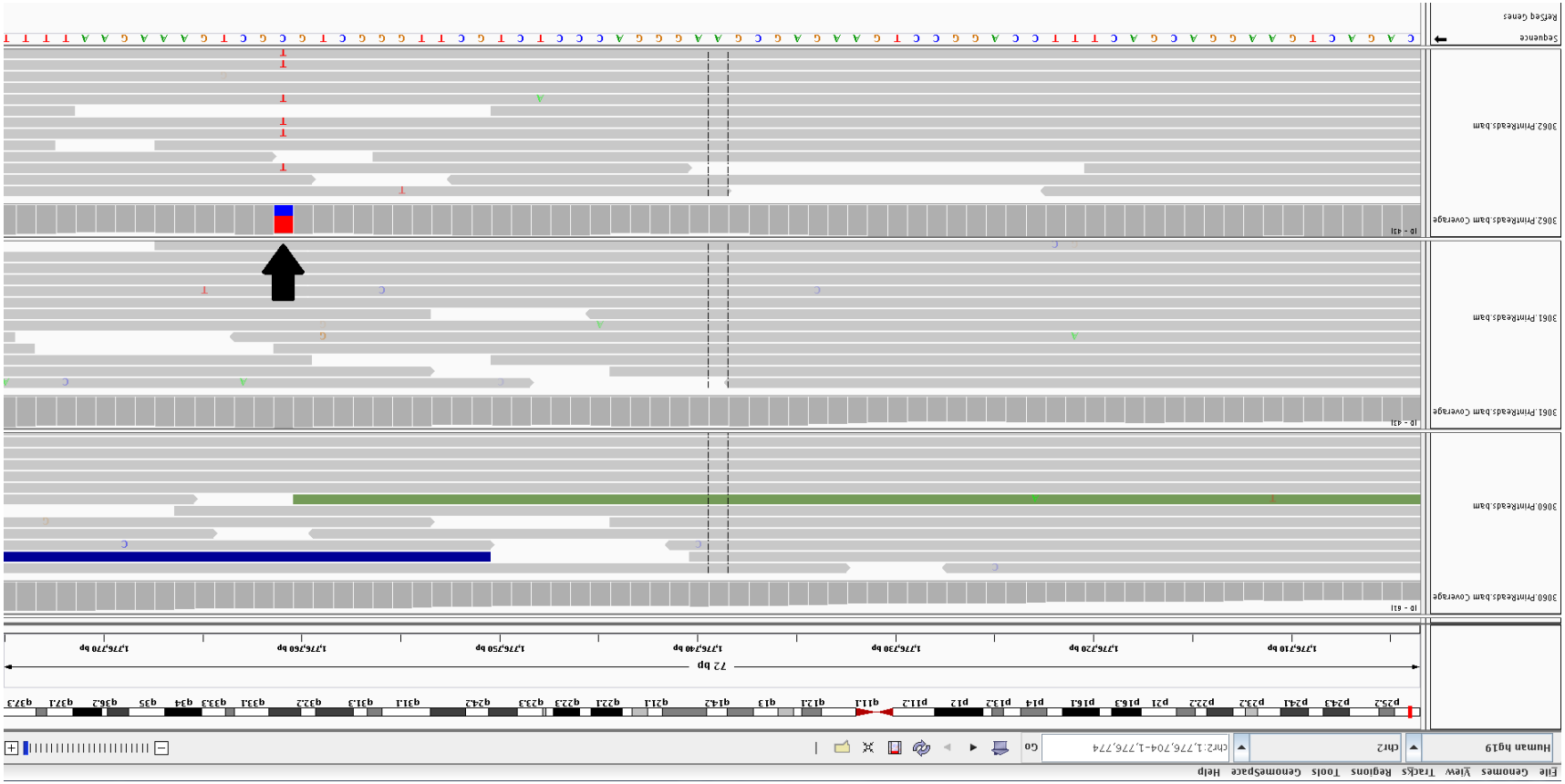

Figure 2-1: A typical screen seen when using IGV, showing a large number of high quality reads and a heterozygous variant (indicated by the arrow). The green and blue reads are reads where the other read of the pair maps to a different chromosome, suggesting unusual alignment. Adapted from IGV browser (Thorvaldsdottir et. al., 2013).

The IGV browser was also used to assess deletions that were found using paired end reads. Paired end reads will normally be aligned close to each other, but in the event of a deletion the pairs of reads are often not able to align correctly, and one of the two is often displaced. When this occurs IGV highlights the reads in red and labels them discordant. This can then be used to find deletions by eye, as if a deletion is suggested by the workflow and there are two well defined blocks of discordant reads, it is a good indication that the deletion is real, while if no or few reads are discordant it suggests that there is no deletion. In addition to discordant reads, IGV showing a drop in coverage to either half of the average number of reads or no reads at all is a good indication of a deletion being present. An example of discordant reads and a drop in coverage indicating a deletion can be seen in figure 2-2.

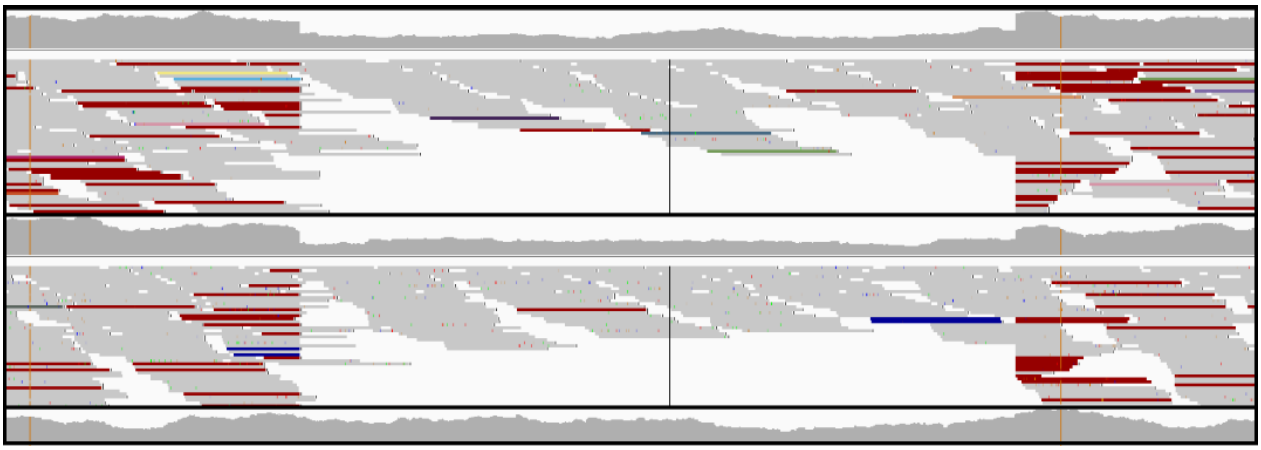

Figure 2-2: An example of a large number of discordant reads and a drop to half of the average coverage in IGV indicating a heterozygous deletion. The discordant reads are red, shown either side of the deletion below the grey coverage chromatogram. Adapted from IGV browser (Thorvaldsdottir, Robinson, & Mesirov, 2013).

2.3.3 Gene function

If a variant was found in a gene that had a function known to be involved with the phenotype shown by the affected individual, as well as meeting the other requirements, it became a strong candidate. On the other hand, if the variant was in a gene known to tolerate loss of function or accrue variants at a high rate, such as the T-cell receptor genes, it was considered less likely to be causal. If a candidate variant was intronic and had no clear significance in causing disease, i.e. was not likely to affect splicing or another regulatory element, it was also excluded as otherwise there would be too many candidate variants remaining to reasonably analyse through a secondary sequencing methodology.

2.3.1 ExAC browser

In addition to gene function, the ExAC prediction of how likely it was that the gene could tolerate a variant that could cause loss of function was used to determine if a variant in a certain gene was a candidate for causing the phenotype seen. The ExAC database defines a variant as causing Loss of Function (LoF) if it is a nonsense, splice acceptor, or splice donor variant caused by single nucleotide changes (Lek et al., 2016). ExAC assigns each gene a score between 0 and 1 based on the number of expected variants of these types compared to the number observed. The closer the LoF intolerance score is to one, the less likely it is that the gene can tolerate loss of function. Any gene with a loss of function tolerance below 0.4 as calculated by ExAC was excluded as unlikely to be a cause of dominant genetic disease. For recessive variants, as the LoF score was not relevant, the presence of homozygous loss of

function variants present in the gene was used instead. If high-confidence homozygous loss of function variants were present, variants in that gene were excluded.

2.3.2 UCSC Genome Browser

For the deletion variants, the start and end positions given in the vcf were entered into the University of California Santa Cruz (USCS) Genome Browser (URL: http://genome.ucsc.edu/). This genome browser contains information on deletions from the Database of Genomic Variants (DGV), which contains information from ExAC and 1000 genomes (Kent et al., 2002). If multiple deletions with similar breakpoints (within 100bp of either end) to the candidate deletions were present then the deletion was deemed too common to be causative.

2.4 HGVS (Human Genome Variation Society) nomenclature

Following vetting of the candidate variants, standardised nomenclature was constructed to describe them using the HGVS system. The HGVS system is comprised of three names. The first describes the genomic position of the variant, the second which transcript contains the variant and its position within said transcript, and third describes the predicted change the variant will cause in the protein produced. If the candidate variant was intronic no protein level name was constructed as the effects of the variant on the protein produced could not be predicted. This nomenclature makes it easy for others to locate the variant in question themselves as well as see the effect it is predicted to have at a glance. The HGVS nomenclature for each candidate variant was constructed following the HGVS guidelines (den Dunnen et al., 2016). In addition, Mutalyzer version 2.0.25 was used to aid in constructing the names (Wildeman, van Ophuizen, den Dunnen, & Taschner, 2008).

2.5 Sequencing for variant confirmation using the Sanger methodology

2.5.1 The Polymerase Chain Reaction

2.5.1.1 Primer Design

Primers flanking a candidate variant were chosen using Primer 3 version 4.0 (Untergasser et al., 2012), before checking for multiple priming sites using the In-Silico PCR tool on the UCSC website (URL https://genome.ucsc.edu/cgi-bin/hgPcr). They were then ordered from Integrated DNA Technologies (IDT) and diluted to 5µM before use.

2.5.1.2 Optimisation

An optimisation reaction was run on a temperature gradient for the primer annealing step between 52°C and 64°C using control human DNA to determine the best temperature to use for the real reaction. The temperature that produced the highest quality PCR product was subsequently used for each reaction. The optimisation program consisted of:

1. Incubation at 94°C for 4 minutes.
2. Incubation at 94°C for 30 seconds.
3. Gradient across the wells from 52°C to 64°C for 30 seconds.
4. Incubation at 72°C for 1 minute.
5. Repeat steps 2 to 5 35 more times and end.

2.5.1.3 Reaction Conditions

Following optimisation samples were run in a PCR machine at the optimal annealing temperature for each reaction, as determined by the optimisation reaction. This program consisted of:

1. Incubation at 94°C for 4 minutes.
2. Incubation at 94°C for 30 seconds.
3. Incubation at optimal annealing temperature for 30 seconds.
4. Incubation at 72°C for 1 minute.
5. Repeat steps 2 to 5 35 more times and end.

2.5.2 Gel electrophoresis

After PCR, the samples were run for 30 minutes at 10V/cm in a 2% agar gel in 0.5XTBE containing 0.5 µg/mL ethidium bromide, before being visualised on a UV transilluminator to determine if amplification had occurred successfully. In addition to the samples, a standard 100bp ladder was used to aid in visualising length. Success was determined based on the presence and clarity of the expected bands, as well as if the length of the DNA amplified matched expectations.

2.5.3 Confirmation by sequencing using the Sanger methodology

2.5.3.1 Pre-sequencing

If gel electrophoresis showed sufficient amplification of the samples for sequencing, a pre-sequencing reaction was used to clean up the PCR products before the sequencing reaction. First, 0.7µL of 5x sequencing buffer (composed of 0.4M Tris-HCL with a pH of 9.0 and 10mM $MgCl_2$), 0.075µL of Exonuclease 1 (Biolabs, 20U/µL), and 0.4µL shrimp alkaline phosphatase (USB, 1U/µL) was added to 2 µL of PCR product. This was then made up to a total volume of 5 µL using distilled water. The reactions were then incubated at 37°C for 15 minutes before a further 15-minute inactivation step at 80°C in a thermal cycler.

2.5.3.2 Sequencing

Following the pre-sequencing reaction, the reactions were made up to a 10µL final volume through the addition of 0.6µL of Big Dye Ready Reaction Mix (ABI), the appropriate sequencing primer (0.16µM final concentration), and the appropriate volume of distilled water. The sequencing primers were either the forward or reverse primer used to amplify the sample during PCR. The reactions then underwent a thermal cycler step consisting of 25 cycles of: 95°C for 10 seconds, 50°C for five seconds, and 60°C for four minutes. Following this the samples were precipitated by adding 62µL precipitation mix (consisting of 77.5% ethanol and 10mM sodium acetate) followed by centrifuging at 4000rpm for 10 minutes in a plate centrifuge with an 18cm radius. The resulting supernatant was removed and the pellet washed with 100µL of 80% ethanol. This was followed by a second centrifuge step of 4000rpm for 10 minutes. The

supernatant was again removed and the pellet was then left to air-dry upside down, with no light exposure for a minimum of 30 minutes.

The samples were then resuspended and sequenced by the Genetic Analysis Services (Department of Anatomy, University of Otago, Dunedin) using a 3730xl DNA Analyzer. The resulting chromatograms were then visualized using FinchTV version 1.4.0 (URL http://www.geospiza.com/Products /finchtv.shtml). Candidate variants were located manually and validated through their presence or absence in the appropriate samples.

3 A case of unexplained OPD1 syndrome

3.1 The OPD1 family

In 1999, a family with a child affected by an unusual case of OPD1 was referred to Professor Stephen Robertson while the gene that caused OPD1 was still unknown (personal communication). The affected individual from this family presented a very severe OPD1 phenotype, with the addition of intellectual impairment, and was therefore clinically diagnosed with severe OPD1. The parents were unaffected (see figure 3-1). Since the time of this patient's referral, variants in *FLNA* have been established as a cause for OPD-1, and in all other families with this diagnosis that have been examined the disorder can be accounted for by variants in this gene. The causal variants are usually small deletions or missense variants that alter the filamin A protein (Robertson et al., 2003). To attempt to determine the variant causing such severe OPD1 in this individual, the exons of the *FLNA* gene had been extensively sequenced using the Sanger methodology, but no variant that could cause disease was found. In addition to sequencing of the *FLNA* gene, array-CGH had been performed, but again no causative variant was discovered.

The failure of the CGH array, as well as sequencing of the F*LNA* gene using the Sanger methodology, to discover a causative variant in this family makes it likely that the causative variant is in a different gene, which would be unknown in the history of OPD1 cases. Finding the variant in this family causing OPD1 may greatly enhance our understanding of the *FLNA* gene

and with what its product interacts, or identify alternative pathways that give rise to this phenotype.

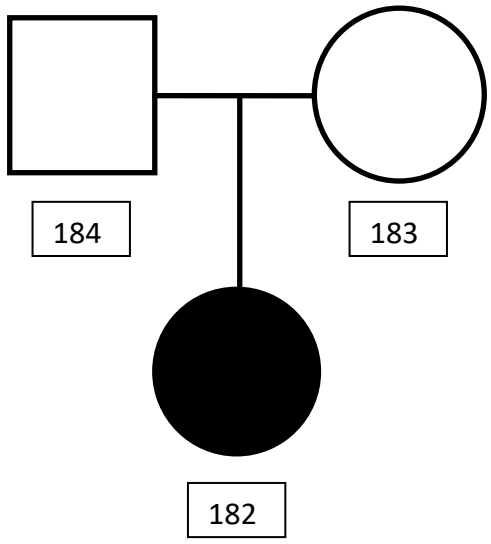

Figure 3-1: A visual representation of the pedigree of the trio affected with OPD1. The affected individual is coloured in black. Circles are female while squares are male. In this family both the mother and father are unaffected while their daughter has OPD1.

3.2 Confirmation of family relationships

First, it was necessary to confirm that the relationships between the three individuals in the trio were as recorded. This was because models based on Mendelian inheritance will be used to attempt to find the disease-causing variant in the affected child. If for some reason, such as the wrong individual being collected or mislabelling of the sample or file, the wrong samples were used in the analysis then the analysis would be invalid.

To confirm that the three individuals were in fact mother, father and child, the number of Mendelian violations at biallelic autosomal variants present in the trio was calculated. There were 238 Mendelian violations found. In order to determine if this was within the acceptable range for a family trio relationship, the process was repeated with an altered pedigree, where

the child from the Larsen family (see section 4.1) was substituted instead of the child from this family. This produced 1390516 Mendelian violations, which showed that the 238 found was within the acceptable range and the relationships were consistent with those recorded.

3.3 Investigation of the *FLNA* gene region

As all previous cases of OPD1 for which a molecular diagnosis has been obtained have been caused by variants in the *FLNA* gene, any variants found in the *FLNA* gene that could be causal would be prime candidates to be causing the disease in the child. Therefore, the *FLNA* gene region was examined to locate any variants in the gene that could be causative. One *de novo* variant was found in the *FLNA* gene in the affected individual, but it was not located in a coding sequence. It also fell within the VQSRTrancheSNP99.90to100.00 truth sensitivity tranche, which suggested that there was a high chance it was a false positive. Upon viewing the variant in IGV it was found that the variant was in an area of low coverage at the end of multiple reads (see Appendix two), and many of the partners of the reads at this location were mapped to other chromosomes, suggesting that there is an issue with mapping to this location. Because of this, it was deemed unlikely to be causing this individual's OPD1.

As there was no candidate variant found in the *FLNA* gene in this individual, it was necessary to examine the rest of the genomic data for variants that could be causal.

3.4 Investigation of a *de novo* variant model for SNVs, short insertions and deletions (short variants)

3.4.1 Filtering for *de novo* variants

Although the inheritance in this family could be interpreted in a number of ways, the simplest hypothesis is a *de novo* variant producing a dominant trait. *De novo* variants are of particular interest because only a small number are expected to occur in a child, so when a disease previously unknown in the family appears at the same time, they become good candidates for investigation as to whether they are causative. Therefore, to determine if any *de novo* variants that could be causative were present in the affected child, the genomic data was filtered to find any locations where the mother and father matched the reference sequence but the child did not, where the variant was likely to have an impact on gene function. In addition to filtering for *de novo* variants, the data was filtered to exclude variants present at an allele frequency of greater than 0.001 in either the ExAC or 1000 Genomes databases. In addition to filtering for rarity, only variants that were annotated as having either a high or moderate impact according to their classification by the SnpEff tool were retained after filtering. Variants classified as such by SnpEff are those that are predicted to change the amino acid sequence of the resulting protein in some way, which makes them good candidates for causing phenotypic changes.

There were five *de novo* variants listed as high impact and 18 as moderate. Of these several appeared to be artefacts and many were excluded for other reasons (see Appendix three). For example, there were two suggested *de novo* variants in the *MTMR1* gene, but they occurred within a microsatellite repeat. As the workflow used to locate candidate variants is known to

have difficulties aligning around repeat elements such as microsatellites, and PCR has issues with slippage when sequencing such areas, these variants were excluded. Another candidate *de novo* variant eliminated was a loss of function variant in the *TRIP10* gene, as *TRIP10* has a tolerance of loss of function score of 0.35 in the ExAC database. This means that the *TRIP10* gene is known to have a high tolerance for loss of function variants and as such this variant was unlikely to be causing the disease seen in the affected child.

Of the three variants not excluded, one listed as high impact was in a splice donor site for exon 20 of the *CREBBP* (CREB Binding protein) gene (HGVS transcript NM_001079846.1:c.3665+1G>C). Variants in the *CREBBP* gene are known to cause Rubinstein-Taybi syndrome (Rubinstein & Taybi, 1963), a disease with some features similar to OPD-1 with the addition of mental retardation. The other two remaining candidates were single base substitutions in *RBMXL3* (HGVS protein NM_001145346.1(*RBMXL3*_i001):p.(His860Arg)) and *USP17L11* (HGVS protein NM_001256854.1(*USP17L11*_i001):p.(Phe29Val)). Both of these variants were listed as moderate impact, and neither of these genes had been previously associated with any disorders. Not much information was available about the function of these genes, which made them hard to assess. Many of the reads covering the *USP17L11* variant had a mapping quality of zero, which meant that there was more than one location in the genome that they were equally able to be mapped to, and one allele was far more common than the other at the variant location.

Both the *CREBBP* and *RBMXL3* variants looked plausible in IGV, as they were not in a repetitive region and had adequate read depth and good allelic balance (see figure 3-2). These variants were therefore considered to be reasonable candidates, but the *CREBBP* variant was considered to be more likely as it was in a known disease gene, where the disease had features similar to OPD1, and the variant would disrupt splicing. Therefore, the area around and including the *CREBBP* variant was sequenced according to the Sanger methodology to confirm its presence.

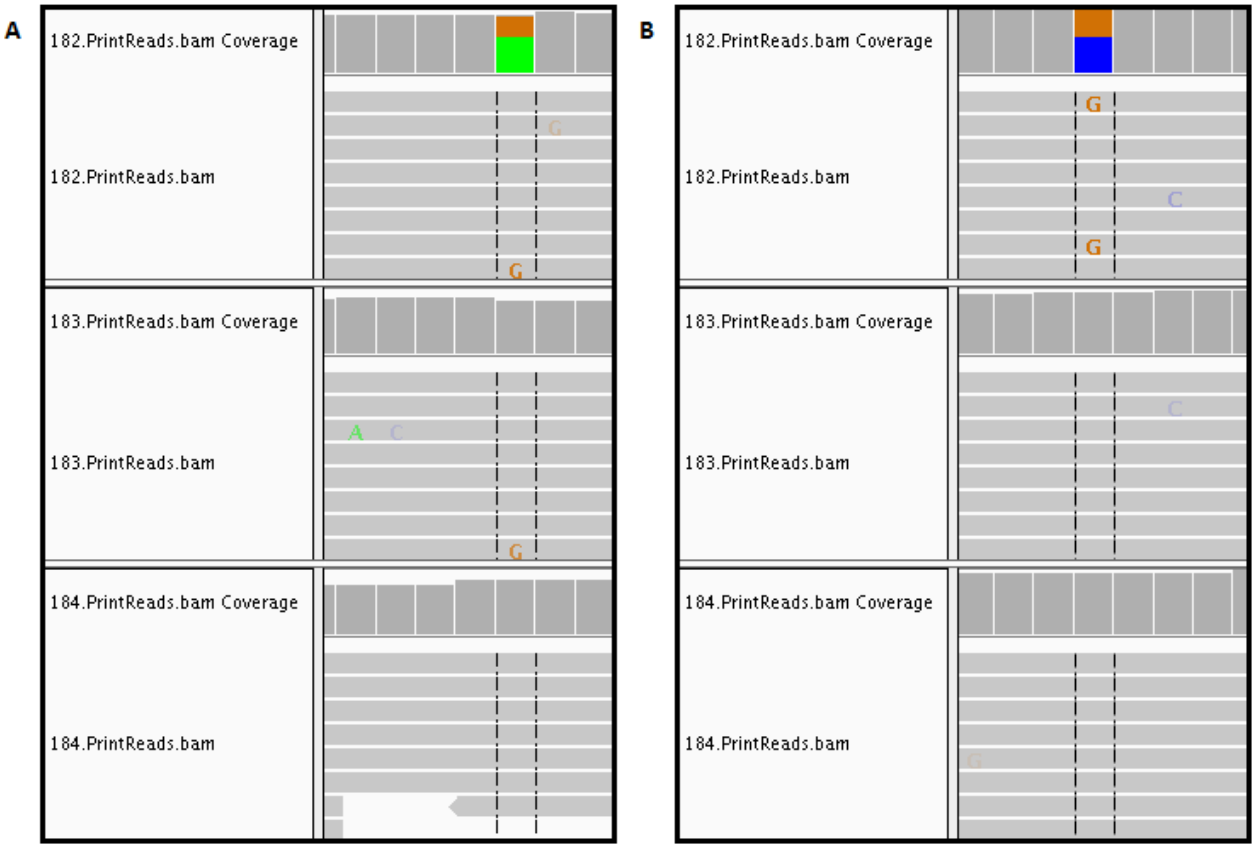

Figure 3-2: The RBMXL3 and CREBBP variants as seen in IGV. A is the RBMXL3 variant, while B shows CREBBP. In both images 182 is the affected child, 183 is the mother and 184 is the father. Adapted from IGV (Thorvaldsdottir et al., 2013).

3.5 Investigation of a *de novo* model for deletions

In addition to searching for short *de novo* variants, larger *de novo* deletions were filtered for using the data generated from GenomeStrIP SV. To locate *de novo* deletions a filter was constructed that found any locations in the genome where there a deletion was present, and the mother and father matched the reference sequence, but the child did not. Deletions that appeared more than once in the control population were also filtered out, and in addition to filtering for rarity, candidates that were not of sufficient quality (i.e. those that did not meet the PASS criteria from GenomeStrip) were filtered out. After filtering, five candidate deletions remained, of which four were not convincing after examination of the coverage in the IGV browser for this trio due to the low number of discordant paired-end reads, high numbers of reads with a mapping quality score of 0, and large numbers of low quality reads surrounding the deletion. The remaining one, however, was 122,284bp long and resulted in the deletion of four entire olfactory receptor genes on chromosome 1 (see figure 3-3). Although *de novo* deletion variants are rare (only four convincing *de novo* deletion variants were found using GenomeStrIP SV in a population of 87 trios for which data was available), it was decided that this deletion was unlikely to be causing the phenotype seen in the affected individual. This decision was made due to the nature of the genes deleted and the fact that a variant very likely to be causative of the phenotype was found in the *CREBBP* gene.

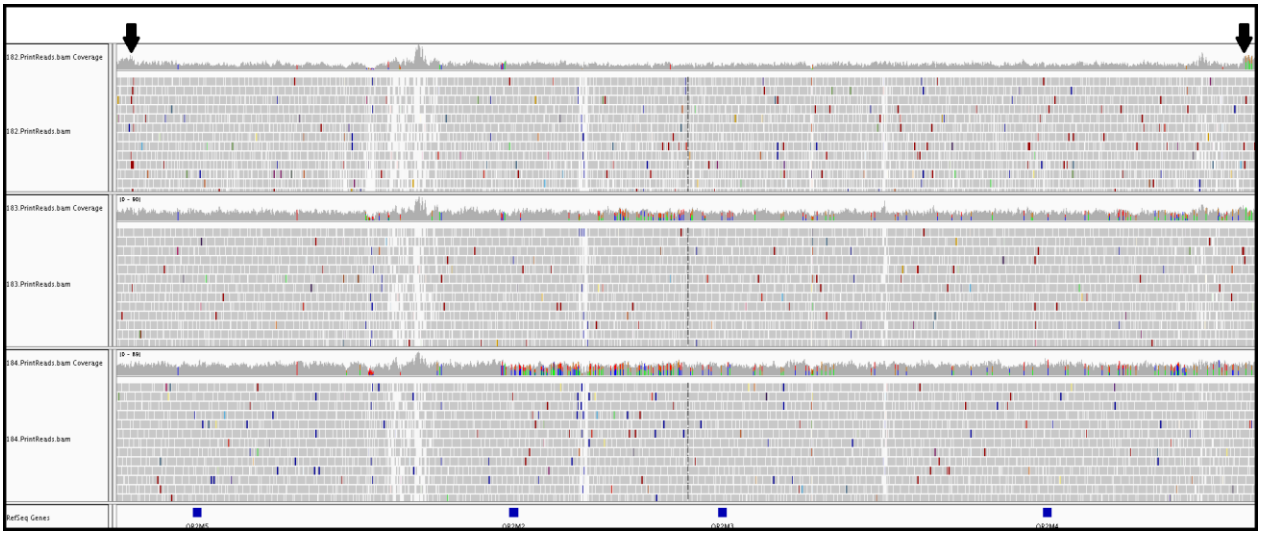

Figure 3-3: An image showing the de novo deletion found in individual 182 that deletes four olfactory receptor genes, as seen in IGV. It is 122,284bp in length. The arrows indicate the approximate breakpoints at either end of the deletion. However, due to the nature of the genes deleted this deletion is unlikely to be causing the phenotype seen in the affected individual.

3.6 Confirmation of *CREBBP de novo* variant

Sequencing across the location of the suspected *CREBBP* variant according to the Sanger methodology confirmed the *CREBBP* variant, and therefore that the affected individual almost certainly has Rubinstein-Taybi syndrome and not OPD-1 (see figure 3-4).

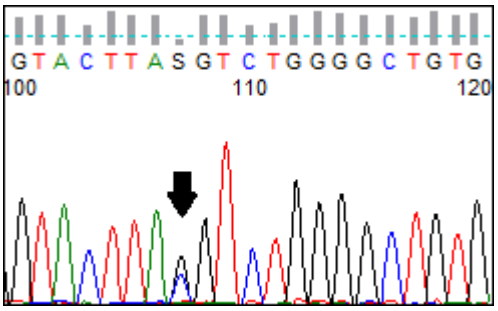

Figure 3-4: The chromatogram produced during validation of the CREBBP variant by sequencing using the Sanger methodology. The variant is indicated by the arrow and is heterozygous. The chromatogram was produced using FinchTV v1.4.0 (http://www.geospiza.com/Products /finchtv.shtml).

The Human Genome Variation Society nomenclature for this variant at the genomic and transcript levels is in Table one. As this variant is a splice donor variant, it is difficult to predict the effects of the variant on this protein. There are two main possibilities; the variant could cause readthrough into the next intron, which may then be translated until a stop codon is encountered, or it is possible that increased skipping of the exon may occur. If the variant causes readthrough into the intron, it is predicted that the protein will truncate almost immediately (HGVS protein NM_001079846.1(*CREBBP*_i001):p.(Thr1223*)). In the event that a premature stop codon was created, it is likely that nonsense mediated decay would occur, and no protein would be produced. The other possibility is that the variant would cause the exon to be skipped with increased frequency. As the length of the exon is a multiple of three bases, skipping it would result in an in-frame deletion in the protein. The HGVS protein nomenclature for this would therefore be NM_001079846.1(*CREBBP*_i001):p.(Tyr1196_Thr1222del).

Even though the exact effects of this variant on the protein are difficult to predict, it is likely that they will cause loss of function.

Table 1: The Human Genome Variation Society nomenclature for the CREBBP variant at the genomic and transcript levels. The protein level name could not be produced, as the exact effect of this variant on the protein could not be predicted. It is likely, however, that the variant will cause loss of function of the protein, as it affects a splice donor site. Tolerance of LoF (loss of function) is the score given by the ExAC database to show how likely the gene is to be able to tolerate a LoF variant, between 0 and 1. A gene with a score of 0 is tolerant of loss of function.

Gene, Chromosome & Base number	**HGVS genomic**	**HGVS transcript**	**HGVS Protein**	**Tolerance of Loss of Function Score**
CREBBP 16: 3801726	NC_000016.9:g.3801726C>G	NM_001079846.1:c.3665+1G>C	Splice donor site affected	1.0

3.7 Discussion

As the variant in *CREBBP* was confirmed by sequencing using the Sanger methodology and predicted to truncate the protein produced, it was concluded that this individual has Rubinstein-Taybi syndrome instead of OPD-1, and that the initial diagnosis was mistaken.

It is possible that the initial diagnosis was incorrect, and remained uncorrected for such a long time, for multiple reasons. Firstly, the case occurred before the OPD spectrum of disorders was well defined and before the techniques used here were developed. In addition to this, the diagnosis was made using second hand information (a description of the affected individual with no images). Finally, OPD1 and Rubinstein-Taybi have some overlap in their phenotypes which can make them difficult to distinguish. For example, individuals affected by either OPD1 or Rubinstein-Taybi often have broad thumbs and generalised bone dysplasia, in addition to short stature.

Although it is slightly disappointing that this is a relatively trivial explanation for the case and it doesn't illuminate any new biology as was initially hoped, this finding provides value to the family concerned in terms of a more confident diagnosis and a more accurate prognosis. For example, individuals affected by Rubinstein-Taybi have been found to be more susceptible to tumour formation and cardiac abnormalities (Miller & Rubinstein, 1995). Knowing about these issues makes it possible for the family and physicians to take steps to control and minimise their effects.

4 A case of unexplained Larsen syndrome

4.1 The Larsen Family

The second family in this investigation also consists of unaffected parents with an affected child (see figure 4-1). The child was diagnosed very early on to have the phenotype typical of Larsen syndrome, with scoliosis and joint dislocations, with the additional trait of weak teeth prone to chipping, which has not previously been described in association with Larsen Syndrome (Robertson, personal communication). The weak teeth may be coincidental, but as unusual traits such as this are often diagnostically useful, it may be a possible clue to assist new gene identification. For all individuals diagnosed with Larsen syndrome for which a molecular diagnosis has been obtained, a causal variant has been found in either the *FLNB* gene (in the case of autosomal dominant Larsen) or the *B3GAT3* gene (autosomal recessive Larsen).

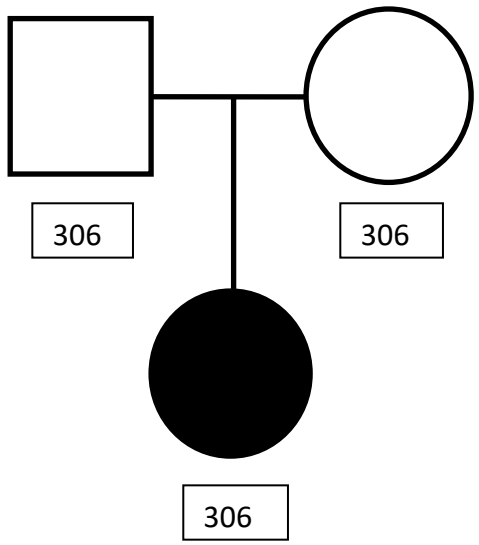

Figure 4-1: A visual representation of the pedigree of the trio affected with Larsen Syndrome. Clear shapes are unaffected individuals while the affected individual is coloured black. Circles are female while squares are male. In this family both the mother and father are unaffected while their daughter has Larsen Syndrome.

However, in the case of this family no causative variant has been found despite a thorough search. Initially the exons and the exon-intron boundaries of *FLNB* had been sequenced individually using the Sanger methodology. In addition, MLPA was performed over the gene, but no causative variant was found (Robertson, personal communication). Following this, array CGH was performed using the Aligent 180k oligoarray platform, but again nothing likely to be causal was found. Finally, in 2014 the exomes of the affected child and their parents were sequenced, but again no convincing variants were found under either dominant or recessive models of analysis.

Since such thorough techniques have been unable to find a cause for the Larsen syndrome in this family, it is likely that, as in the family affected by OPD1, the Larsen is of novel cause. Also, as the exomes of the family have been examined previously it is possible that the cause is outside of the coding region of the genome. This means that finding the variant causing Larsen disease in this family could give us insight into how the filamin genes are regulated or with what they interact, helping us understand the pathway of which they are a part.

4.2 Confirmation of family relationships

As for the OPD1 family, it was necessary to confirm that the relationships between the three individuals in the trio were as recorded. This was because models based on Mendelian inheritance will be used to attempt to find the disease-causing variant in the affected child. If for some reason, such as the wrong individual being collected or mislabelling of the sample or file, the wrong samples were used in the analysis then the analysis would be invalid. To confirm that the three individuals were in fact mother, father and child the number of autosomal biallelic

Mendelian violations present in the trio was calculated. There were 260 Mendelian violations found. In order to determine if this was within the acceptable range for a first-degree relationship, the process was repeated with an altered pedigree, where the parents from the OPD1 family were substituted into the trio from this family. This produced 1390516 Mendelian violations, which showed that the 260 found was within the acceptable range and the relationships were consistent with those recorded.

4.3 Investigation of a *de novo* variant model for SNVs, short insertions and deletions (short variants)

4.3.1 Filtering for *de novo* variants

As for the family affected by OPD1, the simplest hypothesis for what is causing the disease in this family is that of a causal *de novo* variant. As before, the genomic data was filtered for locations where both the mother and father matched the reference sequence but the child did not. Variants present at an allele frequency of greater than 0.001 in either the ExAC or 1000 Genomes databases were excluded. In addition to filtering for rarity, only *de novo* variants were annotated as having either a high or moderate impact were kept.

As variants in the *FLNB* gene have been found to be responsible for causing dominant Larsen syndrome in every family for which a molecular diagnosis has been obtained, any *de novo* variants found in the *FLNB* gene would be excellent candidates to be causative in this family. Therefore, the list of variants that remained after filtering was examined for any variants in the *FLNB* gene that could be causative. There were no *de novo* variants in this gene or in the

surrounding regions of a megabase either side, so *de novo* variants across the rest of the genome were examined. This distance was chosen as it has previously been described that enhancers can be up to a megabase from the gene they affect (Pennacchio et al., 2013).

4.3.2 Remaining candidate *de novo* variants

Across the rest of the genome there were seven *de novo* variants that remained after applying the filter above. Of these, six were deemed unlikely to be causal for any of a number of reasons, such as the gene being likely to tolerate loss of function mutations. For example, one of the suggested *de novo* variants was in the *MUC4* gene, which is known to normally be variable without causing a deleterious phenotype. Another candidate *de novo* variant removed was in the *FAM186A* gene, and was located in a cluster of a large number of common variants (four common variants within five bases of the variant, in addition to one allele being much more common than the other. This unusual clustering of suggested variants and lack of allelic balance made it likely that the candidate variant was an artefact of alignment, so it was excluded. A list of the excluded candidate variants can be found in Appendix four. The remaining candidate *de novo* was in the *ASXL3* gene on chromosome 18 at base 31,319,283 and is a T>C substitution predicted to change a Serine to a Proline in the resulting protein. *ASXL3* has a loss of function tolerance score of 1.0 in ExAC, which means that it is predicted to be highly intolerant of loss of function variants. The HGVS nomenclature and other details for this variant can be found in Table two.

Table 2: Gene name and genomic location for the remaining candidate variants from all different models after filtering as well as their HGVS standardised names. The number of homozygous LoF variants in ExAC shows how many high-confidence LoF variants were present in the ExAC database. If such variants were present it was more likely that loss of function variants in that gene were tolerated without causing disease.

Gene, Chromosome & Base Number	HGVS Genomic	HGVS Transcript	HGVS Protein	Number of homozygous LoF variants in ExAC	Model of origin
ASXL3 18: 31319283	NC_000018.9:g.31319283T>C	NM_030632.1:c.1915T>C	NM_030632.1(*ASXL3*_i001):p.(Ser639Pro)	0	*De novo*
FIGN 2: 164467336	NC_000002.11:g.164467336_164467337insTC	NM_018086.2:c.1005_1006insGA	NM_018086.2(*FIGN*_i001):p.(Tyr336Aspfs*40)	0	Compound heterozygous
FIGN 2: 164467339	NC_000002.11:g.164467340_164467342delTCC	NM_018086.2:c.1000_1002delGGA	NM_018086.2(*FIGN*_i001):p.(Gly334del)	0	Compound heterozygous
PKD1 16: 2143546	NC_000016.9:g.2143546C>T	NM_000296.3:c.11012G>A	NM_000296.3(*PKD1*_i001):p.(Arg3671Gln)	0	Compound heterozygous
PKD1 16: 2159667	NC_000016.9:g.2159667T>C	NM_000296.3:c.5501A>G	NM_000296.3(*PKD1*_i001):p.(Asn1834Ser)	0	Compound heterozygous
SLCO4A1 20: 61291767	NC_000020.10:g.61291768_61291769delGA	NM_016354.3:c.893_894delAG	NM_016354.3(*SLCO4A1*_i001):p.(Glu298Alafs*254)	0	Compound heterozygous
SLCO4A1 20: 61291771	NC_000020.10:g.61291772delT	NM_016354.3:c.896delT	NM_016354.3(*SLCO4A1*_i001):p.(Leu299Argfs*46)	0	Compound heterozygous

4.4 Investigation of recessive models for short variants

4.4.1 Filtering for recessive variants

As the *de novo* candidates discovered in this family were not certain to be causative, it was necessary to continue to search the genome for likely variants. Therefore, a model to filter for recessive variants, both homozygous and compound heterozygous, was constructed.

4.4.1.1 Homozygous variants

In order to find variants where the affected child had inherited a homozygous variant from the parents, the genomic data was filtered to find locations where the affected individual had a homozygous variant and both the mother and the father were heterozygous. In order to narrow

down the list of candidates to a manageable number it was necessary to exclude variants with a genotype quality of less than ten, as such a low quality score indicates low confidence that the base call is accurate . Variants present at an allele frequency of more than 0.001 in either the ExAC or 1000 Genomes databases were also removed. The candidate variants were also filtered on their impact; all variants other than those predicted to have a high or moderate impact were removed.

After this filtering, there were 21 homozygous variants across 15 genes, of which only one remained as a candidate after further examination. For example, three loss of function candidate variants in the *NBPF10* gene were excluded because individuals in the ExAC database have previously been found to have homozygous loss of function variants in that gene, without a severe developmental disorder. This makes it unlikely that the recessive loss of function variants in the *NBPF10* gene in this family are causing the deleterious phenotype seen. Another variant removed after further examination was in the *FIGN* gene, and was removed as it occurred in a microsatellite repeat, which meant that it was probable that the variant was due to either slippage during PCR or an error during the alignment process. The full list of variants and their reasons for exclusion can be found in Appendix five.

4.4.1.2 Compound heterozygous variants

To find variants where the affected child had inherited a heterozygous variant from one parent, in addition to a different heterozygous variant within the same gene from the other parent, the genomic data was filtered to simultaneously select for locations where the child and father were heterozygous but the mother was homozygous reference. Then locations where the child and

mother were heterozygous but the father was homozygous reference were filtered for. Following this a list was made of all the genes where the child had inherited at least one variant from the mother, as well as at least one variant from the father. This list was then further filtered to include only variants that were not present or were present at a frequency of less than 0.001 in the ExAC and 1000 Genomes populations. Variants of low quality were also removed by excluding VQSRTrancheSNP99.90to100.00, as well as by removing all variants with a genotype quality of less than 10.

After filtering, there were 74 compound heterozygous variants across 24 genes, of which two remained as candidates after further examination. These are defined in Table two. The other candidates were excluded for a variety of reasons; for example, three candidate variants remained after filtering in the *NBPF10* gene, but in the ExAC database there are multiple benign homozygous loss of function variants recorded that are different to the candidates found in this analysis, one of which occurred 36 times. This makes it likely that *NBPF10* is able to tolerate homozygous loss of function without causing a deleterious phenotype, so the variants in *NBPF10* were removed from the list of candidates. Other candidates, such as two in *PIK3AP1*, were excluded because they were located in microsatellites, which meant that it was probable that the variant was caused either by slippage during PCR or an error during the alignment process. Another pair of candidates excluded was the pair of variants in the *TRBV10-1* gene. These variants were excluded because the *TRBV10-1* gene is a T-cell receptor gene, which means that it is normally highly variable (Wilson *et. al.*, 1988). The full list of variants and their reasons for exclusion can be found in Appendix five.

4.4.2 Remaining recessive candidates

4.4.2.1 Homozygous candidates

After further examination of the candidate homozygous variants to determine if they could be causal there were none that remained as a candidate for causality.

4.4.2.2 Compound heterozygous candidates

After further examination of the candidate variants that fit the compound heterozygosity model to determine if they could be causal, two pairs remained as candidates that required confirmation. Details of these variants can be found in table two.

4.5 Investigation of a *de novo* model for deletions

As for the OPD1 family, the genomic data of the trio affected by Larsen syndrome was examined for the presence of deletions detected using the GenomeStrIP SV workflow as described in methods section 2.2.3. The deletion data was then filtered to locate variants that fit various inheritance models, the first of which was that of *de novo* deletions responsible for a dominant trait. In order to find *de novo* deletions, locations where the child carried a deletion compared to the reference sequence but both parents did not were found. Following this, variants that were seen more than once in the control population were removed, as they were deemed to be too common. There were no *de novo* deletions found in this family, so a model to find deletions contributing to a recessive trait was constructed.

4.6 Investigation of recessive models for deletions

4.6.1 Filtering for deletions with possible recessive effects

The second model examined using the deletion data was that of deletions with possible recessive effects, including both homozygous and compound heterozygous situations. In the case of compound heterozygous contributions, the hypothesis was that there was an inherited heterozygous deletion within a gene or close enough to one to possibly affect its function from one parent, and the affected gene also contained a short variant inherited from the other parent that could be affecting its function. Combined this could make the gene non-functional. The short variants considered as candidates in combination with a deletion were filtered to only include those predicted to have a high or moderate effect on a protein, even though doing so risked missing a possible causal intronic or intergenic variant. This was necessary due to the large number of variants outside coding regions that would otherwise have remained as candidates, many of which would be impossible to evaluate with regards to their ability to affect gene function. In addition to genes containing a deletion, genes within a megabase of the approximate breakpoint of a heterozygous deletion were included as it was possible that the deletion damaged a regulatory element such as an enhancer.

4.6.1.1 Homozygous deletions

There was one homozygous deletion identified in the affected child, located on chromosome 8 approximately between bases 70,319,872 and 70,325,202. The closest gene to this deletion is *LINC01603*, which is between bases 70,337,106 and 70,360,479 and is a long intergenic non-coding RNA. However, a deletion with predicted breakpoints within 50bp of either end of this

deletion was previously found in five individuals out of 1151 in the 1000 Genomes phase one database (URL http://dgv.tcag.ca/dgv/app/variant?id=esv2673314&ref=hg19). The overlap in location of these deletions suggests that they are the same as the candidate deletion, which made it too common to be considered a candidate for causality of a rare but highly penetrant single gene disorder.

4.6.1.2 Compound heterozygous deletions

There were 21 heterozygous deletions detected in the affected child. None of the deletions contained any exons and none obviously affected splice sites in the NCBI RefSeq transcript database (URL https://www.ncbi.nlm.nih.gov/refseq/). Of the 21 deletions, only one was within one megabase of a gene, in addition to that gene containing a possible variant predicted to affect its function. The gene, called *C18orf61*, was roughly 500,000 bases away from the deletion and contains a short variant predicted to affect splicing.

4.6.2 Remaining recessive deletion candidates

For these reasons, of the 22 deletions found only one was considered a candidate to be causing the disease in this family, in combination with a variant in a gene of unknown function called *C18orf61*. The deletion is roughly 1470 base pairs long between bases 11,730,592 and 11,732,061 on chromosome 18 within an intron of *GNAL* (G protein subunit alpha L) (reference genome GRCh37/hg19). This location is just an estimate as the exact breakpoint of the deletion cannot be accurately determined to within 50bp at this time. The variant that in conjunction with this deletion is a candidate is an intronic A>G substitution predicted to affect a splice donor site at base 12,213,597 in *C18orf61*, ~500,000bp from the deletion. *C18orf61* is an

uncharacterised open reading frame (ORF) of unknown function located on chromosome 18 between bases 12200778 and 12224710 of transcript NC_000018.9 (reference genome GRCh37/hg19). Because nothing is known about the function of *C18orf61* or how it could be affected by the deletion in question, it is difficult to evaluate this variant. This means that although the variant in *C18orf61,* in conjunction with the intronic *GNAL* deletion, remains a candidate it is of low priority and was not sequence validated.

4.7 Confirmation of the remaining candidates

The candidates that remained after both filtering and further examination were then validated by sequencing using the Sanger methodology. Of those, several were found to have been false positives and were therefore eliminated as candidates, including a pair of compound heterozygous variants in the *SCLO4A1* gene. However, one in the *ASXL3* gene was visible in the sequence data and therefore validated, and so remains as a viable candidate for causing the disease in this family (see figure 4-2).

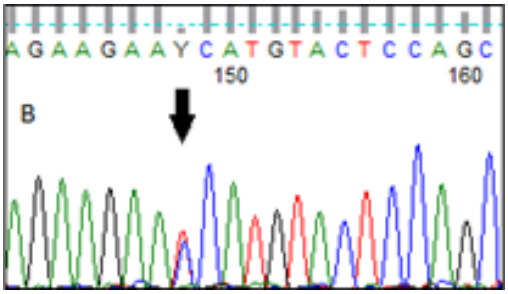

Figure 4-2: The chromatogram produced during validation of the ASXL3 variant by sequencing using the Sanger methodology. The variant is indicated by the arrow and is a heterozygous de novo variant. These chromatograms were produced using FinchTV v1.4.0 (http://www.geospiza.com/Products /finchtv.shtml).

4.8 The remaining candidates

4.8.1 ASXL3

ASXL3 (Additional Sex Combs-Like 3) is found on chromosome 18 and is predicted to be a transcriptional regulator that epigenetically regulates some homeotic genes throughout development (Katoh, 2015). In the affected individual, the *de novo* variant is in the penultimate exon of the *ASXL3* gene. Truncating variants in this exon have previously been described as causing Bainbridge-Ropers syndrome, one of which was only 60bp away (Bainbridge et al., 2013). Of the four variants described by Bainbridge et al. all caused early termination of the protein and therefore likely made the protein non-functional. Bainbridge-Ropers syndrome has a phenotype that could not be confused with Larsen syndrome (Robertson, personal communication), including psychomotor retardation, feeding problems and ulnar deviation of the hands.

The variant in this family is not truncating, but changes a polar amino acid (Serine) for a non-polar one (Proline), possibly altering protein structure and therefore function. This could lead to the phenotype seen in the affected individual, in addition to explaining why the phenotype is different. As *ASXL3* regulates transcription during development it is possible that altered function of the protein it produces could affect the expression of other genes.

4.8.2 *C18orf61* in conjunction with the intronic *GNAL* deletion

C18orf61 is an uncharacterised open reading frame (ORF) of unknown function located on chromosome 18 between bases 12200778 and 12224710. As an ORF is simply a stretch of

sequence that does not contain stop codons, more information is required to confirm the existence of a protein coding gene, but ORFs of this length are good indicators of possible protein-coding regions (Deonier, 2005). This means that variants within it could cause a change in phenotype. However, as there is no known function for *C18orf61* and any there is a lack of information about possible regulatory effects of the intronic *GNAL* deletion, it is difficult to evaluate this variant, so although it remains a candidate it is of low priority and was not sequence validated.

4.9 Discussion

As the remaining candidates do not make a compelling case for causality in terms of their known function, it is not possible to confidently determine the exact causal variant in the family affected by Larsen syndrome. In order to accomplish this in the future, it would be useful to recruit additional atypical Larsen families with no causative variant in either the *FLNB* or *B3GAT3* gene. Such families can then be used to find variants affecting genes in common, which provides support for any novel genes suggested to be causal. If this is not possible, as atypical Larsen families seem to be very rare, the causal variant could theoretically be determined through a variety of functional studies using one of a number of experimental systems, depending on the proposed function of the gene containing the variant suggested to be causal. For example, cell based assays of the variant could be performed, or the affected gene could be knocked out in zebrafish (*Danio rerio*) or mice (*Mus musculus)* if an ortholog exists. Performing such functional studies might allow predictions to be made about the phenotypic consequences of such variants in humans.

It is possible that the real causal variant was accidentally removed during filtering in this investigation. Such an outcome could occur for a number of reasons, the most likely of which is that the causal variant was in a non-coding region and excluded due to a lack of the knowledge required to evaluate them. It could also have been removed due to having low base quality calls or issues during the annotation phase that caused it to be removed by the filters. It is also possible that the real causal variant could have been missed because it was not in the aligned data available. For example, the area the variant is in could be an area of the genome that doesn't sequence well and is therefore not available in the aligned dataset, such as an area rich in GC bases. The causative variant could also have been missed because it was in an uncompleted part of the genome, as there are still sections of the human genome that are not fully known (Mostovoy et al., 2016).

Additionally, the workflow that discovered and annotated structural variants could have missed a causal deletion. This is because it is less sensitive than the workflow that annotates SNVs and indels, and so requires stringent filtering to remove false positives, which can also remove real variants. There is also the possibility that the causal variant in this family is a different type of structural variant such as a duplication, which would not have been detected. In order to find such variants, it may be possible to use the GenomeStrIP CNV workflow (Handsaker et al., 2015) or other tools which can detect duplications and copy number variants as well as deletions. In order to find other structural rearrangements, such as inversions and translocations, other tools such as BreakDancer (Chen et al., 2009) can be used.

Although no concrete answer was found as to the cause of the Larsen syndrome seen in this unusual family, one of the remaining candidates may be the causal variant, and in the event that

another atypical Larsen family is discovered, the *ASXL3* gene and *C18orf61* should be of particular interest and thoroughly searched. If another variant that could be causal is found in either *ASXL3* or *C18orf61* in another family, then the case for it causing the phenotype seen is made much stronger, such that functional studies could be recommended.

5 Discussion

5.1 The OPD1 family

5.1.1 How did a misdiagnosis happen?

The *de novo* variant in the *CREBBP* (CREB Binding Protein) gene was validated by sequencing using the Sanger methodology, and was predicted to truncate the protein produced by destroying a splice site. Many types of variant in the *CREBBP* gene are associated with various diseases including Rubinstein-Taybi syndrome (Petrij et al., 1995), non-Hodgkin lymphoma (Pasqualucci et al., 2011) and acute myelogenous leukaemia (Petrij et al., 1995). Rubinstein-Taybi is most often caused by sporadic heterozygous *de novo* variants and has a phenotype similar to that of OPD1 with the addition of intellectual impairment (Rubinstein & Taybi, 1963). As variants similar to the one found in this family have been shown to cause Rubinstein-Taybi syndrome it was concluded that this individual has Rubinstein-Taybi syndrome instead of OPD-1, and that the initial diagnosis was mistaken. This is supported by the phenotype of the affected individual because the intellectual impairment shown is not typical of OPD1 but is common for Rubinstein-Taybi.

There could be several reasons for the initial diagnosis being incorrect and remaining uncorrected for such a long time. Firstly, OPD1 and Rubinstein-Taybi have a lot of overlap in their phenotypes which can make them difficult to distinguish. For example, individuals affected by either OPD1 or Rubinstein-Taybi often have broad thumbs and generalised bone dysplasia, in

addition to short stature. In addition to this, there are a huge number of rare genetic syndromes that often have complicated and subtle phenotypes that can vary between individuals.

Secondly, the case presented in 1999, before the OPD spectrum of disorders was well defined. This meant that the clinicians diagnosing such individuals were less able to confidently diagnose someone as one of the OPD disorders. In addition to this, as the OPD disorders are quite rare it would not be unusual for the clinician in question to have never encountered someone affected by one of them before. These factors combined could have led to the clinician involved with this family incorrectly described the child as having OPD1. In addition to this, the diagnosis was accepted here using second hand information (a description of the affected individual with no images), which makes it more likely that the clinician will make a mistake during diagnosis.

Finally, the techniques used here and the knowledge necessary to apply them were not yet developed at the time. The human genome project had not yet been completed and the function of many genes now known had not yet been discovered, which meant that clinical diagnosis was almost always carried out based on phenotype alone. This meant that if two rare syndromes had a similar phenotype, and had not been encountered by the clinician involved before, then it was possible for the diagnosis to be incorrect. If this occurred then the techniques used today were not able to be used to distinguish between the two possible syndromes, as at the time of this case it was very expensive to use genetic techniques to diagnose someone. Even if it could be afforded, genetic diagnosis was unlikely to be considered due to the lack of knowledge necessary for it to be useful.

For these reasons it is not surprising that genetic syndromes are frequently misdiagnosed. However, a misdiagnosis can often be corrected by molecular analysis of the affected individual, as such an analysis can identify the exact variant likely to be causing the disease and therefore which gene is functioning abnormally to give the phenotype seen. If abnormal function of that gene is linked to another phenotypically similar syndrome that matches what is seen in the patient, as was the case with this family, then it is likely that the patient has that disease and the misdiagnosis can be corrected. The ability to correct such misdiagnoses is one of the reasons that molecular approaches such as whole genome analysis are valuable to clinicians.

5.1.2 Rubinstein-Taybi Syndrome

Rubinstein-Taybi is an autosomal dominant disease with a phenotype similar to that of OPD-1 including broad thumbs, generalised bone dysplasia and short stature (Rubinstein & Taybi, 1963). In addition to these traits, those affected by Rubinstein-Taybi normally have mental impairment, whereas those affected by OPD1 usually do not. The phenotype of the affected individual from the family involved in this investigation is in hindsight typical of Rubinstein-Taybi, strengthening the hypothesis that the *CREBBP* variant is causal and the individual was initially misdiagnosed. In addition to the traits listed, individuals affected by Rubinstein-Taybi have been found to be more susceptible to tumour formation, particularly neoplasms of the nervous system (Miller & Rubinstein, 1995). Individuals affected by Rubinstein-Taybi are also likely to have heart problems, as in previous literature cardiac abnormalities were found in 32.6% of those studied that had the disease (Stevens & Bhakta, 1995).

5.1.3 Outcomes

Although this is a relatively trivial explanation for the case and it doesn't illuminate any new biology as was hoped, this finding provides value to the family concerned in terms of a more confident diagnosis and a more accurate prognosis. As those affected by Rubinstein-Taybi are known to have a higher incidence of cardiac abnormalities and have a higher rate of tumour formation than those with a normal phenotype, correcting this diagnosis means that the family will be aware of these risks and better able to deal with them should they arise in the future.

In addition to correcting the misdiagnosis, this finding also strengthens the hypothesis that OPD1 is exclusively caused by gain-of-function variants in the *FLNA* gene, as no variants known to cause OPD1 have yet been found in any other gene, even in this family where it initially seemed likely that that was the case. However, it remains possible that OPD1 can be caused by variants in another gene that simply haven't been discovered yet.

Finally, this case is a good example of the value of having access to high quality clinical information, such as images of the affected individual. Because the initial diagnosis was made based on only second-hand information with no images available, it was incorrect and remained uncorrected for nearly 20 years. While in this instance the prognosis for Rubinstein-Taybi was similar to that of OPD1, at other times this may not be the case, and an incorrect diagnosis could lead to long term harm for the individual. For example, many of the disorders similar to OPD1 share some traits, but have different long-term outcomes, so high quality information is needed to accurately differentiate them. In particular, access to images of the affected

individual showing the abnormal traits is very useful in the event that an in-person examination is not possible by the clinician.

5.2 The Larsen Family

5.2.1 The remaining candidates

The remaining candidate variants are in genes that are not known to have a function that could cause Larsen syndrome if disrupted. Because of this, it is not possible to definitively state that either of them is the exact cause of this individual's Larsen syndrome. However, it is possible that one of them is causal. This can be confirmed in the future if another family with unexplained Larsen is discovered, as the *ASXL3* gene and *C18orf61* can be investigated to find any variants that could be causal.

5.2.1.1 *ASXL3*

One of the remaining candidates is a *de novo* variant in the *ASXL3* gene, which is known to cause Bainbridge-Ropers syndrome if damaged. *ASXL3* produces an epigenetic scaffold protein and has a loss of function tolerance score of 1.0 in ExAC, which means that it is highly intolerant of single base variants that could cause loss of function, to such an extent that the gene is haploinsufficient (Lek et al., 2016). In addition to this, the candidate *de novo* is predicted to change an amino acid in the protein produced by SnpEff 4.2, and therefore has the potential to alter function. The candidate variant has also not been found previously. These factors make it a reasonable candidate, but not one strong enough that it can be deemed causal without further supporting evidence.

5.2.1.1.1 Gene function and known associated diseases

ASXL3 produces a protein that is an epigenetic scaffold, involved in the regulation of nuclear receptor-dependent transcription as well as repressing EZH2-dependent transcription (Katoh,

2015). It is part of a gene family that is critical to development, as components of the signal transduction pathway from members of the ASXL gene family affect the pathways of other epigenetic regulators during physiological, pathological and developmental processes (Katoh, 2015). Dysregulation of epigenetic processes such as this can cause a variety of diseases in humans because epigenetic regulators are critical for normal cellular differentiation (Ordovas & Smith, 2010).

Because of this, variants that affect the function of ASXL genes can cause a variety of diseases in humans, including Bohring-Opitz syndrome (*ASXL1*) and many cancers such as breast, liver and bladder cancer (Katoh, 2015). Truncating variants in *ASXL3* in particular have been implicated in autism (De Rubeis et al., 2014), Bainbridge-Ropers syndrome (Bainbridge et al., 2013), and melanoma (Berger et al., 2012). The variants previously described to cause Bainbridge-Ropers syndrome are located throughout the gene with the exception of the beginning. As the last two exons of *ASXL3* are much larger than the others, most of these variants are in the same exon as the variant found in this family, exon 11 (Kuechler et al., 2017), one of which was only 60bp away (Bainbridge et al., 2013). However, Bainbridge-Ropers syndrome has a phenotype that could not be confused with Larsen syndrome, including psychomotor retardation, feeding problems and ulnar deviation of the hands (Robertson, personal communication).

5.2.1.1.2 The candidate variant

The exon containing the candidate variant in this family is the second to last one in the *ASXL3* gene, and all previously described disease-causing variants in this exon truncate the protein

produced, either making it non-functional or possibly causing a gain of function, leading to Bainbridge-Ropers syndrome. The candidate variant in this family is not truncating and therefore is less likely to cause Bainbridge-Ropers syndrome as it is a different kind of variant to those that classically cause the disease. However, it is possible that it could cause a different phenotype such as Larsen's syndrome, or be completely benign. The missense variant seen in this family, particularly because it is predicted to affect protein structure, could cause a gain of function or non-function of the protein produced. Specifically, the variant is predicted to change a Serine, a polar amino acid, to a non-polar one, Proline.

5.2.1.1.3 Possible effects of the variant

The variant in this family changes a polar amino acid (Serine) for a non-polar one (Proline), possibly disrupting protein structure and therefore function. This could affect the function of the protein as usually polar amino acids are located on the outside of the protein to allow for interaction with other molecules while non-polar ones are in the interior (Alberts et al., 2014). Changing the charge of an amino acid can therefore change the structure of the protein and how it interacts, which could possibly lead to gain of function and the phenotype seen in the affected individual. As *ASXL3* epigenetically regulates transcription during development, either non-function of the protein it produces or a change in its function could affect the expression of other genes. In addition to this, the fact that the variant is not truncating but still could affect protein function would explain why the phenotype is different when compared to other diseases associated with the *ASXL3* gene.

5.2.1.1.4 What does this mean for Filamin B?

In the event that the *ASXL3 de novo* candidate variant is found to be causal through future work, it might suggest that filamin B interacts with *ASXL3* in some way. For example, it was found to be causal it may be suggested that transcription of the *FLNB* gene is normally epigenetically regulated by *ASXL3*, and that *FLNB* is sensitive to altered function of that protein. This would mean that either too much or too little filamin B was produced during development due to loss of normal transcriptional regulation of *FLNB*, which could then lead to the disorder seen in this family.

If this was the case it would enhance our understanding of filamin B and how it operates during development, including what molecules or receptors with which it is likely to interact. This would in turn lead to a better understanding of the cellular processes involved, especially for the development of the skeleton during development.

5.2.1.2 *C18orf61* in conjunction with the intronic *GNAL* deletion

The second candidate that remains is a heterozygous single base variant in *C18orf61*, an open reading frame (ORF), in conjunction with a heterozygous intronic deletion in the *GNAL* gene, roughly 500,000bp away. Both the function of *C18orf61* and how it could be regulated by the *GNAL* deletion are unknown, and so it is very difficult to assess the possible impact this candidate could have. Although the two variants in combination remain a candidate for causing the disease in this family, due to the lack of information available about how they could cause the phenotype seen they are of low priority.

5.2.2 What does this mean for the family and others affected by Larsen Syndrome?

Even though the exact causal variant in this family could not be determined as it was for the OPD1 family, this study still provides some value as two candidate variants remained, one in the *ASXL3* gene and the other in *C18orf61* in combination with an intronic deletion in the *GNAL* gene between bases 11,730,592 and 11,732,061 on chromosome 18. However, neither of them is a compelling candidate. The *ASXL3* gene is not known to have any function that when disrupted could lead to the phenotype seen, and *C18orf61* is both of unknown function and it is not known how the intronic *GNAL* deletion could affect its regulation. It is possible that the true causal variant was missed for any of a number of reasons. For example, it is possible that the disease seen in this individual is not caused by genetic factors but environmental ones, or the causal variant could be in a region not adequately covered for technical reasons such as repetitive regions and regions high in GC bases. It is also possible that the causal variant lies in a non-coding region, which could mean that although the sequence data was available it was not adequately screened due to the large number of variants present and the current lack of knowledge about the functional effects of non-coding variants. In addition to these reasons, it is possible that the causal variant could have been missed in this analysis because it is in a region of structural polymorphism that is not adequately described by the reference sequence used and so was not visible, or because the cause of the disease is a rare structural variant not covered by this analysis such as a duplication.

For the reasons above, the remaining candidates are not compelling enough to justify undertaking functional investigation of them. However, in the event of the discovery of another family affected by this rare type of Larsen syndrome it would be possible to strengthen the case for either variant being causal if the gene in question was found to harbour a significant variant in the second family. In this event, there is the possibility of further functional work being undertaken to determine if either variant is causal, including cell based assays of the candidate variant or knockdowns of the affected gene in zebrafish (*Danio rerio*) to observe the resulting phenotypes. In the event that the causal variant is identified through such means it would be useful for physicians, who would be better able to diagnose the disease through genetic techniques. This in turn can provide value to the affected family, as it can lead to better management strategies and an improved diagnosis.

5.2.3 How could we determine the causal variant in the Larsen family?

Although neither of the remaining candidates are compelling enough to justify functional assessment at this point, in the future evidence obtained from another individual or family may strengthen the case for one variant to be causal, to the point that functional testing of the variant will be necessary. If this occurs there are several options that can be used to determine if the candidate thought to be causal is, including cellular assays based on the function of the proteins thought to be affected and knockdowns in animal models.

5.2.3.1 Cellular assays

One way a strong candidate can be assessed functionally is to cause the suggested causal variant in an appropriate cell line, which can then be assessed using criteria relevant to the

expected phenotype. For example, as the product of the *ASXL3* gene is involved in the epigenetic regulation of many genes during development, it would be necessary to choose a cell line that expresses the *ASXL3* gene as well as one or more of the genes it is known to regulate, in a suitable environment for normal regulatory function. If this was achieved it would then be possible to determine what effect changing the *ASXL3* gene by causing the candidate variant would have on that pathway and therefore whether the *ASXL3* variant could be causal.

If the cellular assay showed phenotypic changes in the cell line that could be attributed to the variant introduced, it would support the hypothesis that that variant was causal, and further functional work could be undertaken using an animal model to further investigate the likely effects of the variant in question.

5.2.3.2 Knockdowns in Zebrafish or knockout mice

Following a successful assessment of the candidate variant using cellular assays, or as an alternative, it would be possible to further strengthen the argument for that variant being causal by recreating the candidate variant in an animal model and observing the resultant phenotypic effects in an attempt to recapitulate the relevant parts of the disease phenotype. There are several animal models that could be used, but the most common ones are zebrafish and mice. As zebrafish are cheaper and easier to use than mice, they are more often used to determine if strong candidate variants are in fact causal. However, as mice are more closely related to humans it is more likely that an ortholog will exist in mice than in zebrafish. For example, there is no known ortholog for the *ASXL3* gene in zebrafish but there is in mice. It is also possible that the variant cannot be caused in one or the other model species because it is

lethal. For these reasons, if functional work were ever to be performed using an animal model, it may be necessary to use mice to investigate the variant in the *ASXL3* gene, but zebrafish to examine other variants found by further analysis due to the difference in cost and ease of utilisation.

In addition to the above issues, as these species are biologically different to humans the phenotype produced may not be the same as it would be in humans, which can make it difficult to properly assess the effect of a candidate variant. For Larsen's syndrome, however, if the disease phenotype was replicated in either mice or zebrafish, it could be relatively easy to assess as it consists of mainly bone malformations and deafness, both of which could be assessed in either species even though the exact changes may not be identical to the phenotype in humans. For example, whether the variant caused deafness in Zebrafish or not could be assessed through playing a loud noise and observing the response via a camera. If the Zebrafish did not exhibit the classic startle response then it is likely that their hearing has been impaired, possibly by the introduced variant (Yang *et al.*, 2017). Such functional tests are not currently recommended however, as the support for the remaining candidates is not strong enough to justify them.

5.2.3.3 Finding another family

The most straightforward and likely way the correct causal variant in the Larsen family could be determined is using another family with the same type of unexplained Larsen syndrome. Having a second family containing an affected individual means that their genomic sequence can be compared to those of the current family to find any candidates in the same gene, thereby

greatly enhancing the power of the methods used here and increasing support for implicated genes if they are found to contain multiple independent variants. However, although this way would effectively increase support for a suggested causal gene, it relies on finding another family with the same type of Larsen syndrome as the one used here. Since the current family is the only known case of Larsen syndrome not caused by a variant in either *FLNB* or *B3GAT3*, it can be considered extremely rare, and therefore it is not likely that a second family will be found in the near future.

5.3 Whole Genome Sequencing

5.3.1 Advantages

There are several advantages to using whole genome sequence to locate causal variants compared to exome sequence, namely that whole genome sequence gives higher quality sequence of the exome and provides a greater potential for researchers to find structural variants. One possible advantage of using whole genome sequence is that when using recently developed methods to obtain whole genome sequence PCR-amplification of a library of sequences is not necessary, and there is no capturing or hybridization step (Belkadi et al., 2015). This means that the areas of the genome missed when an exome is produced are more likely to be included in the resulting sequence, as well as reducing the bias towards certain areas of the genome and giving more uniform depth of coverage across the genome in general and therefore more reliable base calls. However, in most cases, including this analysis, PCR is still used to amplify a library of sequences, as the PCR free approach is not standard.

As well as giving more complete coverage of the genome and more reliable base calls than exome data, using whole genome sequence provides greater potential for the discovery of some structural variants, which is more difficult when using exome sequence because there is no simple way to search or filter for them as the available methods are both less accurate and less sensitive.

In addition to these advantages, genome data is becoming easier to utilise all the time. Currently there are many excellent tools available for analysis, especially those available in the Genome Analysis Toolkit (GATK) which includes many of the tools used here. Having these tools available means that if genome sequence data is available, it can be effectively searched for causative variants with relative ease. This means that although in the past searching an entire genome for a candidate might have been incredibly difficult, it can now be done relatively easily. Furthermore, the tools available are improved and new tools are created all the time, which means that in the future using whole genome sequence will become even easier.

A further possible advantage of using whole genome sequence to search for causal variants is that it makes it possible to locate causal non-coding variants. Although exome sequence will contain information on splice site variants, as well as roughly the first 100bp of each intron, it does not contain information about the remaining intronic regions, which can lead to the causal variant not being found. However, the number of non-coding variants is extremely large when using whole genome sequence and there is currently little information available that can be used to effectively filter and assess them. This means that in most cases, the majority of intronic variants are filtered out when using whole genome sequence, so while having sequence data on intronic variants is potentially valuable it is currently not possible to effectively utilise it. Despite

this difficulty, it is expected that in the future as there is further development of databases describing functional elements in non-coding regions, such as the Encyclopaedia of DNA Elements (ENCODE), it will be possible to better examine non-coding variants, providing a further advantage to using whole genome sequence.

5.3.2 Limitations

5.3.2.1 Structural variants

5.3.2.1.1 Finding structural variants is difficult

Despite the increased availability and ease of utilisation of genomic data there are still some issues that indicate a need for improvement. Firstly, even though the tools involved are useful, it is still difficult to reliably find structural variants. Variants that are insertions or duplications are particularly hard to find, which means that if the causal variant is one of these, it can be easily missed. This can lead to other candidates being investigated, which leads to a lot of wasted time and resources. However, as stated above, new and improved tools are constantly being developed that make it easier to find such variants. Additionally, this difficulty with structural variants is not unique to whole genome sequence but is shared by other methods, particularly exome data.

5.3.2.1.2 Population size when looking at structural variants

When looking for structural variants, the large online databases of exome and genome data cannot be used to directly annotate the variants discovered, as the start and end positions of the structural variations they contain are approximate. This means that the structural variants will be seen as being different due to the inherent ambiguity in their start and end points. This

ambiguity makes it necessary to use individuals processed together throughout the genotyping phase of the SV workflow as the population. While this enables a certain level of confidence in calculating the frequency of the variants involved, it is not as useful as using the databases due to the smaller number of individuals involved making it possible that a relatively common variant simply is not represented.

5.3.2.1.3 GenomeStrIP CNV Workflow

The workflow that can be used to locate further structural variants that could be causal is called GenomeStrIP CNV and utilises tools from the GATK (Handsaker et al., 2015). This tool builds on the SV version of this workflow by adding functionality that enables the location of insertions and duplications to be determined and annotated in addition to those of deletions. This new workflow will make it far easier to find structural variants in genomic data, but does not work with exome data. The CNV workflow scans overlapping regions of the genome, analysing read depth to look for evidence of structural variants, as variants such as insertions, deletions and duplications will alter the read depth, such as by decreasing it by roughly half in the case of a heterozygous deletion. However, as it is highly complex and difficult to implement, it was not used for this analysis. In addition, there are many other tools available that are designed to identify structural variants form sequence data, but due to time and difficulty constraints not all of them have been investigated or implemented.

5.3.2.2 Such massive quantities of data are difficult to properly investigate

Another issue that can arise when using whole genomes is that the amount of data involved is gigantic (there are approximately 3 billion base pairs in the human genome). When using whole

genome sequence, the dataset usually contains a number of variants that greatly exceeds what can reasonably be assessed manually. Such amounts of data make it necessary to filter out many of the suggested variants so that it is possible to examine the rest at all. Although this filtering is helpful, as it removes most of the false positive results and leaves only the candidates that are likely to be affecting gene function, it is possible to accidentally filter out the causal variant in the process. Again, however, this issue does not only affect whole genome data but also affects whole exome data, although it is much more pronounced when using genome data. To combat this issue, it is necessary to use conservative filters that retain more of the suggested variants, to remove most of the false positives but as few of the real ones as possible.

5.3.2.3 The reference sequence is an approximation

Whole genome sequence is obtained by aligning fragments of sequence data to a reference sequence. However, it is normal for there to be multiple alleles at many loci in the human genome, and such alternative alleles are often not represented in the reference sequence. This means that it is possible for something to be annotated as a possibly causative variant when in fact it is simply a normal alternative allele. Because of this, it is necessary to ensure that any candidate variants are not simply alleles different to those of the reference sequence by checking the frequency of the supposed variant against a large population.

5.3.2.4 Many false positives are present in the dataset

Finally, it is necessary to note that the sequence data itself contains many false positive variants, which are variants that are annotated as not matching the reference sequence but often do on closer inspection, and so are not candidate variants at all. Such false positives are common

when using genome data and if they are retained after filtering can cause incorrect conclusions to be drawn, so it is necessary to take extra steps to ensure the variants exist. This can include using the Sanger methodology to sequence across the variant's location to prove that it is in fact there.

5.3.3 Is secondary sequencing to detect false positive variants still necessary?

Until recently, it was thought necessary to use a secondary sequencing methodology, such as the Sanger methodology, to strengthen the case that the variants thought to be causing disease were not false positives. This was because high throughput sequencing technology early on often made errors in base calls, either due to low fluorophore concentration or other associated issues, which led to a relatively high number of false positive variants being defined as causing disease (Bell et al., 2011). However, as these sequencing technologies have developed they have become more reliable, so using a secondary sequencing methodology to confirm the presence of variants found by the initial method is becoming less common.

Despite this, in this study, several variants passed the filters applied and so were considered candidates, but when sequencing using the Sanger methodology was used to attempt to verify their presence the resulting sequence was found to match the reference sequence. This suggested that those variants were incorrectly called during initial sequencing, and that the real base matched the reference sequence. If the secondary confirmation sequencing step was not carried out, it is possible that these variants would still be candidates, which would be incorrect. Therefore, it is suggested that using a secondary sequencing method known to provide very

accurate base calls such as the Sanger methodology or PCR followed by high depth Illumina sequencing is still necessary even though current sequencing methods generally lead to accurate base calls. This is because in the event that a base is initially called incorrectly, and because of that an incorrect variant is ascribed causality, a great deal of time and resources could be wasted trying to understand how the gene involved interacts with other factors to produce the phenotype seen.

However, even though a second sequencing methodology returning the same base at a given position as the first method strengthens the base call, it is also possible that the second method is also incorrect. For example, the Sanger methodology has low sensitivity when multiple alleles are present, such as in the case of mosaicism, which can lead to incorrectly classifying candidate variants as false positives. In addition, PCR followed by high depth Illumina sequencing is not very economical when confirming the accuracy of only one possible variant, although this can be resolved by running multiple fragments down the same lane to confirm multiple variants at the same time.

5.3.3.1 Will it remain necessary to use secondary sequencing to detect false positive variants?

Although it is still necessary to use a secondary sequencing methodology to confirm the presence of many candidate variants, the advent of several new sequencing technologies that give highly accurate base calls across incredibly long reads are likely to both negate the need for a reference sequence and provide an alternative to short read sequencing methodologies such as Illumina sequencing. Two of these technologies, PacBio SMRT sequencing and Oxford

Nanopore sequencing, provide incredibly long reads of 10,000bp on average from single molecules (Vembar et al., 2016). Producing such long reads greatly improves the mappability, or uniqueness, of reads within a genome (Sims, Sudbery, Ilott, Heger, & Ponting, 2014), which makes it possible to unambiguously align the sequences and therefore assemble genomes without the use of a reference genome, i.e. *de novo*. In addition to this, long reads can span repetitive elements and complex regions without introducing errors that lead to false positive results. Finally, using long reads to assemble a genome enables the determination of structural variants, as at such lengths it is possible to sequence across most of them (Merker et al., 2016).

In addition, although the rate at which these methods make incorrect base calls is relatively high, the base calling errors are random rather than systematic, which means that a relatively low read depth is capable of minimising the number of incorrect calls in the final sequence data (Roberts, Carneiro, & Schatz, 2013). This in turn greatly reduces the number of false positive variants produced, which means that it is generally unnecessary to confirm the existence of candidate variants through further sequencing methods.

However, despite these advantages, it is often not practical to use this methodology because it is very expensive as well as being incredibly computationally demanding. As such, it is not yet used routinely on the scale of whole human genomes and so was not appropriate to use for this analysis. In the future, it is expected that utilising such sequencing technologies will become less expensive and more accessible, and therefore an appropriate choice for the determination of causal variants in Mendelian disease.

5.4 Future analysis

While no further work is recommended at present, it is possible that in the future a second family affected by unexplained Larsen syndrome will be discovered. In this event, the *ASXL3* gene and *C18orf61* should be of particular interest, as if variants that could be causal are found in these genes in a second family the case for it causing the Larsen phenotype seen is made much stronger, to the point that functional studies would be recommended to determine if the gene in question was truly causal. In addition, it may be worth revisiting the Larsen family case in several years, once better tools for locating variants that are currently hard to detect, such as structural variants, are available, as a causal variant that was missed in this analysis could possibly be located.

6 References

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014). Molecular Biology of the Cell: Garland Science

Alkuraya, F. S. (2016). Discovery of mutations for Mendelian disorders. Hum Genet, 135(6), 615-623. doi:10.1007/s00439-016-1664-8

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74. doi:10.1038/nature15393

Bainbridge, M. N., Hu, H., Muzny, D. M., Musante, L., Lupski, J. R., Graham, B. H., . . . Ropers, H. H. (2013). *De novo* truncating mutations in *ASXL3* are associated with a novel clinical phenotype with similarities to Bohring-Opitz syndrome. Genome Med, 5(2), 11. doi:10.1186/gm415

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., . . . Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A, 112(17), 5473-5478. doi:10.1073/pnas.1418631112

Bell, J. (2004). Predicting disease using genomics. Nature, 429(6990), 453-456. doi:10.1038/nature02624

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., . . . Kingsmore, S. F. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med, 3(65), 65ra64. doi:10.1126/scitranslmed.3001756

Berger, M. F., Hodis, E., Heffernan, T. P., Deribe, Y. L., Lawrence, M. S., Protopopov, A., . . . Garraway, L. A. (2012). Melanoma genome sequencing reveals frequent *PREX2* mutations. Nature, 485(7399), 502-506. doi:10.1038/nature11071

Bicknell, L. S., Farrington-Rock, C., Shafeghati, Y., Rump, P., Alanay, Y., Alembik, Y., . . . Robertson, S. P. (2007). A molecular and clinical study of Larsen syndrome caused by mutations in *FLNB*. J Med Genet, 44(2), 89-98. doi:10.1136/jmg.2006.043687

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet, 14(10), 681-691. doi:10.1038/nrg3555

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet, 39(7 Suppl), S16-21. doi:10.1038/ng2028

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., . . . Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods, 6(9), 677-681. doi:10.1038/nmeth.1363

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., . . . Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A, 106(45), 19096-19101. doi:10.1073/pnas.0910672106

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., . . . Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet, 97(2), 199-215. doi:10.1016/j.ajhg.2015.06.009

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., . . . Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature, 515(7526), 209-215. doi:10.1038/nature13772

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., . . . Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat, 37(6), 564-569. doi:10.1002/humu.22981

Deonier, R. C. (2005). Computational Genome Analysis: An Introduction. In M. S. Waterman, S. Tavaré, & SpringerLink (Eds.), Computational Genome Analysis: An Introduction: Springer New York.

Dozmorov, M. G., Adrianto, I., Giles, C. B., Glass, E., Glenn, S. B., Montgomery, C., . . . Wren, J. D. (2015). Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinformatics, 16 Suppl 13, S10. doi:10.1186/1471-2105-16-s13-s10

Dudding, B. A., Gorlin, R.J. & Langer, L.O. (1967). The Oto-palato-digital SyndromeA New Symptom-Complex Consisting of Deafness, Dwarfism, Cleft Palate, Characteristic Facies, and a Generalized Bone Dysplasia. The American Journal of Diseases of Children, 113, 214-221.

Feinleib, M., Garrison, R. J., Fabsitz, R., Christian, J. C., Hrubec, Z., Borhani, N. O., . . . Wagner, J. O. (1977). The NHLBI twin study of cardiovascular disease risk factors: methodology and summary of results. Am J Epidemiol, 106(4), 284-285.

Feng, Y., & Walsh, C. A. (2004). The many faces of filamin: a versatile molecular scaffold for cell motility and signalling. Nat Cell Biol, 6(11), 1034-1038. doi:10.1038/ncb1104-1034

Gorlin, J. B., Yamin, R., Egan, S., Stewart, M., Stossel, T. P., Kwiatkowski, D. J., & Hartwig, J. H. (1990). Human endothelial actin-binding protein (ABP-280, nonmuscle filamin): a molecular leaf spring. J Cell Biol, 111(3), 1089-1105.

Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet, 43(3), 269-276. doi:10.1038/ng.768

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. Nat Genet, 47(3), 296-303. doi:10.1038/ng.3200

Hartwig, J. H., & Stossel, T. P. (1975). Isolation and properties of actin, myosin, and a new actinbinding protein in rabbit alveolar macrophages. J Biol Chem, 250(14), 5696-5705.

Katoh, M. (2015). Functional proteomics of the epigenetic regulators *ASXL1*, *ASXL2* and *ASXL3*: a convergence of proteomics and epigenetics for translational medicine. Expert Rev Proteomics, 12(3), 317-328. doi:10.1586/14789450.2015.1033409

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome Res, 12(6), 996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002

Kuechler, A., Czeschik, J. C., Graf, E., Grasshoff, U., Huffmeier, U., Busa, T., . . . Wieczorek, D. (2017). Bainbridge-Ropers syndrome caused by loss-of-function variants in *ASXL3*: a recognizable condition. Eur J Hum Genet, 25(2), 183-191. doi:10.1038/ejhg.2016.165

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol, 15(6), R84. doi:10.1186/gb-2014-15-6-r84

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), 285-291. doi:10.1038/nature19057

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Lian, G., Dettenhofer, M., Lu, J., Downing, M., Chenn, A., Wong, T., & Sheen, V. (2016). Filamin A- and formin 2-dependent endocytosis regulates proliferation via the canonical Wnt pathway. Development, 143(23), 4509-4520. doi:10.1242/dev.139295

Madan, B., Walker, M. P., Young, R., Quick, L., Orgel, K. A., Ryan, M., . . . Major, M. B. (2016). USP6 oncogene promotes Wnt signaling by deubiquitylating Frizzleds. Proc Natl Acad Sci U S A, 113(21), E2945-2954. doi:10.1073/pnas.1605691113

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753. doi:10.1038/nature08494

Merker, J., Wenger, A. M., Sneddon, T., Grove, M., Waggott, D., Utiramerur, S., . . . Ashley, E. A. (2016). Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease. bioRxiv. doi:10.1101/090985

Miller, R. W., & Rubinstein, J. H. (1995). Tumors in Rubinstein-Taybi syndrome. Am J Med Genet, 56(1), 112-115. doi:10.1002/ajmg.1320560125

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., . . . Kwok, P. Y. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. Nat Methods, 13(7), 587-590. doi:10.1038/nmeth.3865

Oliveira, A. M., Hsi, B. L., Weremowicz, S., Rosenberg, A. E., Dal Cin, P., Joseph, N., . . . Fletcher, J. A. (2004). USP6 (Tre2) fusion oncogenes in aneurysmal bone cyst. Cancer Res, 64(6), 1920-1923.

Ordovas, J. M., & Smith, C. E. (2010). Epigenetics and cardiovascular disease. Nat Rev Cardiol, 7(9), 510-519. doi:10.1038/nrcardio.2010.104

Pasqualucci, L., Dominguez-Sola, D., Chiarenza, A., Fabbri, G., Grunn, A., Trifonov, V., . . . Dalla-Favera, R. (2011). Inactivating mutations of acetyltransferase genes in B-cell lymphoma. Nature, 471(7337), 189-195. doi:10.1038/nature09730

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. Nat Rev Genet, 14(4), 288-295. doi:10.1038/nrg3458

Petrij, F., Giles, R. H., Dauwerse, H. G., Saris, J. J., Hennekam, R. C., Masuno, M., . . . et al. (1995). Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. Nature, 376(6538), 348-351. doi:10.1038/376348a0

Robertson, S. P. (2005). Filamin A: phenotypic diversity. Curr Opin Genet Dev, 15(3), 301-307. doi:10.1016/j.gde.2005.04.001

Robertson, S. P. (2007). Otopalatodigital syndrome spectrum disorders: otopalatodigital syndrome types 1 and 2, frontometaphyseal dysplasia and Melnick-Needles syndrome. Eur J Hum Genet, 15(1), 3-9. doi:10.1038/sj.ejhg.5201654

Robertson, S. P., Twigg, S. R., Sutherland-Smith, A. J., Biancalana, V., Gorlin, R. J., Horn, D., . . . Wilkie, A. O. (2003). Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. Nat Genet, 33(4), 487-491. doi:10.1038/ng1119

Rubinstein, J. H., & Taybi, H. (1963). Broad thumbs and toes and facial abnormalities. A possible mental retardation syndrome. Am J Dis Child, 105, 588-608.

Saleheen, D., Natarajan, P., Armean, I. M., Zhao, W., Rasheed, A., Khetarpal, S. A., . . . Kathiresan, S. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature, 544(7649), 235-239. doi:10.1038/nature22034http://www.nature.com/nature/journal/v544/n7649/abs/nature22034.html#supplementary-information

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res, 30(12), e57.

SenGupta, D. J., & Cookson, B. T. (2010). SeqSharp: A general approach for improving cycle-sequencing that facilitates a robust one-step combined amplification and sequencing method. J Mol Diagn, 12(3), 272-277. doi:10.2353/jmoldx.2010.090134

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet, 15(2), 121-132. doi:10.1038/nrg3642

Stevens, C. A., & Bhakta, M. G. (1995). Cardiac abnormalities in the Rubinstein-Taybi syndrome. Am J Med Genet, 59(3), 346-348. doi:10.1002/ajmg.1320590313

Steward, C. A., Parker, A. P. J., Minassian, B. A., Sisodiya, S. M., Frankish, A., & Harrow, J. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. Genome Med, 9(1), 49. doi:10.1186/s13073-017-0441-1

Stossel, T. P., Condeelis, J., Cooley, L., Hartwig, J. H., Noegel, A., Schleicher, M., & Shapiro, S. S. (2001). Filamins as integrators of cell mechanics and signalling. Nat Rev Mol Cell Biol, 2(2), 138-145. doi:10.1038/35052082

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform, 14(2), 178-192. doi:10.1093/bib/bbs017

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Res, 40(15), e115. doi:10.1093/nar/gks596

Veltman, J. A., & Brunner, H. G. (2012). *De novo* mutations in human genetic disease. Nat Rev Genet, 13(8), 565-575. doi:10.1038/nrg3241

Vembar, S. S., Seetin, M., Lambert, C., Nattestad, M., Schatz, M. C., Baybayan, P., . . . Smith, M. L. (2016). Complete telomere-to-telomere *de novo* assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. DNA Res, 23(4), 339-351. doi:10.1093/dnares/dsw022

Wildeman, M., van Ophuizen, E., den Dunnen, J. T., & Taschner, P. E. (2008). Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat, 29(1), 6-13. doi:10.1002/humu.20654

Wilson, R. K., Lai, E., Concannon, P., Barth, R. K., & Hood, L. E. (1988). Structure, organization and polymorphism of murine and human T-cell receptor alpha and beta chain gene families. Immunol Rev, 101, 149-172.

Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R. H., & Meijer, G. A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res, 34(2), 445-450. doi:10.1093/nar/gkj456

Zhang, H., Zhang, H., Zhang, Y., Ng, S. S., Ren, F., Wang, Y., . . . Chang, Z. (2010). Dishevelled-DEP domain interacting protein (DDIP) inhibits Wnt signaling by promoting TCF4 degradation and disrupting the TCF4/beta-catenin complex. Cell Signal, 22(11), 1753-1760. doi:10.1016/j.cellsig.2010.06.016

7 Appendices

7.1 Appendix One: Depth of coverage for Genomic sequence data

Depth of coverage statistics for each individual were obtained from the bam files using the GATK DepthOfCoverage tool, shown in the table below.

sample_id	mean	%_bases_above_10	%_bases_above_20	%_bases_above_30
182	38.23	98.9	97.9	84.2
183	35.59	98.9	97.4	75.4
184	38.07	99.6	95.5	82.2
3060	45.95	99	98.3	95.1
3061	33.68	98.9	96.5	66.3
3062	35.39	99.5	94.3	73.5

A description of the SNPs and Indels was obtained from variant call file using the Picard CollectVariantCallingMetrics tool, shown in the table below.

Individual	**182**	**183**	**184**	**3060**	**3061**	**3062**
Het/Hom var. ratio	1.809578	1.865308	1.76736	1.900659	1.757828	1.83762
SNPS	3717915	3735288	3707833	3710758	3664367	3704703
Novel SNPS	81655	81918	79936	75818	74884	75836
Filtered SNPS	322552	344343	337087	359224	374762	377973
% in DBSNP	0.978037	0.978069	0.978441	0.979568	0.979564	0.97953
DBSNP TITV	2.084109	2.084736	2.08324	2.085693	2.087563	2.088468
Novel TITV	1.665328	1.661749	1.654358	1.682683	1.659327	1.669248
Indels	505864	554657	548479	511416	559221	569170
Novel Indels	60939	91105	88672	65205	109038	113317
Filtered Indels	4578	5142	5204	5776	5472	5300
% DBSNP Indels	0.879535	0.835745	0.838331	0.872501	0.805018	0.800908
DBSNP Ins/Del Ratio	0.902316	0.895104	0.895744	0.900121	0.895827	0.894879
Novel Ins/Del ratio	0.756015	0.791994	0.789077	0.748311	0.828116	0.825485
Multiallelic SNPs	76409	77007	76421	79806	77014	76669
Multiallelic SNPs in DBSNP	59246	59811	59422	59944	58779	58833
Complex InDels	492632	493043	485218	540368	487867	484805
Complex InDels in DBSNP	418427	418717	413695	452498	416641	414141

7.2 Appendix Two: A low quality *FLNA* variant as seen in IGV

A screenshot of IGV showing the low quality of the *de novo* variant found in the *FLNA* gene in individual 182 of the OPD1 family. Its location is indicated by the arrow. The copy number is much lower than the average and it is situated at the end of many reads. Coloured reads are discordant pairs of reads where one read of the pair has mapped to this location and the other has mapped to a different chromosome. This suggests that there were issues with aligning the pairs correctly and that the variant seen here is unlikely to be real. In addition, the extended stretch of A bases suggests that this region is subject to polymerase stutter during library preparation and cluster generation, which makes it more likely that this variant is an artefact.

7.3 Appendix Three: *De novo* variants in the OPD1 family

A list of variants found through the *de novo* model in the family affected by OPD1 that remained after filtering for any locations where both the mother and father matched the reference sequence but the child did not, where the variant was likely to have an impact on gene function. Any gene with a loss of function tolerance below 0.4 as calculated by ExAC (URL http://exac.broadinstitute.org/) was excluded where the variant was a single base change as unlikely to be causing disease. NA means that there was no information available for loss of function tolerance in ExAC for that gene. In addition, any variant inside a mono or di-nucleotide repeat was excluded as they were likely the result of PCR slippage or an error of alignment. Variants that fell within VQSRTrancheSNP99.90to100.00 were also excluded as they did not pass truth sensitivity.

Gene, Chromosome & Base No	**HGVS Genomic**	**HGVS Transcript**	**HGVS Protein**	**Tolerance of Loss of Function**	**Reason for exclusion**
C2orf27B 2:132552884	NC_000002.11:g.132552884G>C	NM_214461.1:c.482C>G	NM_214461.1(*C2orf27B*_i001):p.(Ala161Gly)	0.01	Gene below LoF cut-off. Does not pass truth sensitivity.
USP17L11 4:9217209	NC_000004.11:g.9217209G>A	NM_001256854.1:c.79G>A	NM_001256854.1(*USP17L11*_i001):p.(Ala27Thr)	NA	Does not pass truth sensitivity.
USP17L11 4:9217215	NC_000004.11:g.9217215T>G	NM_001256854.1:c.85T>G	NM_001256854.1(*USP17L11*_i001):p.(Phe29Val)	NA	
C4orf40 4:71024425	NC_000004.11:g.71024426_71024487delCCCTGTTGCAGCTGAGCCTGCTGCAGAGGCACCTGTTGGAGCTGAGCCTGCTGCAGAGGCA	NM_214711.3:c.458_518delCTGTTGCAGCTGAGCCTGCTGCAGAGGCACCTGTTGGAGCTGAGCCTGCTGCAGAGGCACC	NM_214711.3(*PRR27*_i001):p.(Pro153Leufs*51)	0	Gene below LoF cut-off.
AC104057.1 7:65112677	NC_000007.13:g.65112677G>C	NR_038378.1:n.1208C>G	NA	NA	Does not pass truth sensitivity.
MUC17 7:100676313	NC_000007.13:g.100676313G>A	NM_001040105.1:c.1616G>A	NM_001040105.1(*MUC17*_i001):p.(Ser539Asn)	NA	Does not pass truth sensitivity.

AHNAK2 14:105409235	NC_000014.8:g.105409235A>G	NM_138420.2:c.12553T>C	NM_138420.2(*AHNAK2*_i001):p.(Ser4185Pro)	NA	Does not pass truth sensitivity.
AHNAK2 14:105417326	NC_000014.8:g.105417326T>C	NM_138420.2:c.4462A>G	NM_138420.2(*AHNAK2*_i001):p.(Thr1488Ala)	NA	Does not pass truth sensitivity.
CREBBP 16:3801726	NC_000016.9:g.3801726C>G	NM_001079846.1:c.3665+1G>C	(INTRONIC)	1.00	
TRIP10 19:6746076	NC_000019.9:g.6746076G>A	NM_004240.2:c.985-387G>A	(INTRONIC)	0.35	Gene below LoF cut-off.
SRRM5 19:44117603	NC_000019.9:g.44117603T>A	NM_001145641.1:c.1330T>A	NM_001145641.1(*SRRM5*_i001):p.(Tyr444Asn)	0	Gene below LoF cut-off.
STX16-NPEPL1 20:57266459	NC_000020.10:g.57266459A>T	NM_001204872.1:c.-41-321A>T	(INTRONIC)	NA	Does not pass truth sensitivity.
STX16-NPEPL1 20:57266460	NC_000020.10:g.57266460G>T	NM_001204872.1:c.-41-320G>T	(INTRONIC)	NA	Does not pass truth sensitivity.
CSF2RA X:1422192	NC_000023.10:g.1422192A>G	NM_001161531.1:c.947-624A>G	(INTRONIC)	0	Gene below LoF cut-off. Does not pass truth sensitivity.
TBC1D8B X:106092450	NC_000023.10:g.106092451_106092452delT	NM_198881.1:c.1838-24_1838-23delT	(INTRONIC)	0	Gene below LoF cut-off.
ARMCX4 X:100749038	NC_000023.10:g.100749038C>T	NM_001256155.1:c.5462C>T	NM_001256155.1(*ARMCX4*_i001):p.(Ala1821Val)	NA	Does not pass truth sensitivity.
ARMCX4 X:100749041	NC_000023.10:g.100749041A>G	NM_001256155.1:c.5465A>G	NM_001256155.1(*ARMCX4*_i001):p.(Glu1822Gly)	NA	Does not pass truth sensitivity.
RBMXL3 X:114426583	NC_000023.10:g.114426583A>G	NM_001145346.1:c.2579A>G	NM_001145346.1(*RBMXL3*_i001):p.(His860Arg)	NA	
RBMX X:135956571	NC_000023.10:g.135956571_135956573insGG	NM_001164803.1:c.540+672_540+674insCC	(INTRONIC)	0.95	Does not pass truth sensitivity.
RBMX X:135956575	NC_000023.10:g.135956575G>A	NM_001164803.1:c.540+670C>T	(INTRONIC)	0.95	Does not pass truth sensitivity.
MAGEC1 X:140993884	NC_000023.10:g.140993885_140993991del	NM_005462.3:c.695_801del	NM_005462.3(*MAGEC1*_i001):p.(Pro232Leufs*6)	NA	Contained within an upstream deletion that was filtered out.
MTMR1 X:149867145	NC_000023.10:g.149867150_149867152delCG	NM_003828.2:c.147-518_147-516delCG	(INTRONIC)	NA	In a dinucleotide repeat.
MTMR1 X:149867147	NC_000023.10:g.149867147_149867158delGCGCGCGCACA	NM_003828.2:c.147-521_147-510delGCGCGCGCACA	(INTRONIC)	0.98	In a microsatellite repeat.

7.4 Appendix Four: *De novo* variants in the Larsen family

A list of variants found through the *de novo* model in the family affected by Larsen syndrome that remained after filtering to find any locations where the mother and father matched the reference sequence but the child did not, where the variant was likely to have an impact on gene function. Any gene with a loss of function tolerance below 0.4 as calculated by ExAC (URL http://exac.broadinstitute.org/) was excluded where the variant was a single base change as unlikely to be causing disease (an NA means that there was no information available about LoF in ExAC for that gene). Also excluded were variants in hypothetical transcripts, such as those in *GOLGA6L19*, as there is no direct evidence that they exist. In addition, any variant inside a mono or di-nucleotide repeat was excluded as they were likely the result of PCR slippage or an error of alignment.

Gene, Chromosome & Base number	**HGVS Genomic**	**HGVS Transcript**	**HGVS Protein**	**Tolerance of Loss of Function**	**Reason for exclusion**
MUC4 3:195513537	NC_000003.11:g.195513538_195513539insGG	NM_018406.4:c.4911_4912insCC	NM_018406.4(*MUC4*_i001):p.(Ala1638Profs*1239)	NA	Mucin genes are normally highly variable and contain repeat motifs, which makes them subject to misalignment giving high rates of false positives.
FAM186A 12:50745861	NC_000012.11:g.50745862_50745863insGA	NM_001145475.1:c.4751_4752insTC	NM_001145475.1(*FAM186A*_i001):p.(Gln1585Argfs*71)	NA	In a cluster of common changes, so it is likely this is an artefact of alignment.
GOLGA6L19 15:83014000	NC_000015.9:g.83014000T>C	XM_005272450.1:c.583A>G	XM_005272450.1(*LOC101927601*_i001):p.(Arg195Gly)	NA	XM means hypothetical transcript. Predicted but no direct evidence it exists.
USP6 17:5036226	NC_000017.10:g.5036226C>T	NM_004505.2:c.217C>T	NM_004505.2(*USP6*_i001):p.(Arg73*)	0	
ASXL3 18:31319283	NC_000018.9:g.31319283T>C	NM_030632.1:c.1915T>C	NM_030632.1(*ASXL3*_i001):p.(Ser639Pro)	1	

OAZ1 19:2271510	NC_000019.9:g.2271510C>T	NM_004152.2:c.272C>T	NM_004152.2(*OAZ1*_i001):p.?	0.09	Gene below LoF cut-off.
FAM47A X:34148805	NC_000023.10:g.34148807_34148845delGAGACTGGACGTCCGACGAGTCTTGGGAGGCTCCGAGCG	NM_203408.2:c.1552_1590delCGCTCGGAGCCTCCCAAGACTCGTCGGACGTCCAGTCTC	NM_203408.2(*FAM47A*_i001):p.(Arg518_Leu530del)	0.75	Variant in ExAC at a frequency of 0.00122, too high to be causing this disease.

7.5 Appendix Five: Recessive candidates from the Larsen family

A list of the recessive candidates, both homozygous and compound heterozygous, in the family affected by Larsen syndrome that remained after filtering. In order to find homozygous candidates, the data was filtered for locations where the affected individual was homozygous variant and both the mother and the father were heterozygous, where the genotype quality was greater than ten, the variant was present at a frequency below 0.001 in both ExAC and 1000 genomes, and the variant was predicted to have either a high or moderate impact on the protein produced.

In order to find compound heterozygous candidates, the data was first filtered for locations where the child and father were heterozygous but the mother was homozygous reference. Then locations where the child and mother were heterozygous but the father was homozygous reference were filtered for. Following this a list was made of all the genes where the child had inherited at least one variant from the mother, as well as at least one variant from the father. The candidates were then further filtered to include only variants present at a frequency of less than 0.001 in both ExAC and 1000 Genomes, that were not in VQSRTrancheSNP99.90to100.00, and had a genotype quality greater than 10.

Any gene with high-confidence LOF homozygote variants present in ExAC (URL http://exac.broadinstitute.org/) was excluded where the variant was a single base change as likely to be tolerant of such variants without causing disease. Also excluded were any variants in pseudogenes (*SPATA31E3P*) and those in T-Cell receptor genes (*TRBV10-1*) as they are normally highly variable without causing disease phenotypes (Wilson, Lai, Concannon, Barth, & Hood,

1988). In addition, any variant inside a mono or di-nucleotide repeat was excluded as they were likely the result of either PCR slippage or an error during alignment. Variants that fell into VQSRTrancheSNP99.90to100.00 were also excluded.

As there were too many candidate variants remaining to reasonably analyse through secondary sequencing those that were intronic with no clear significance in causing disease were excluded.

Compound Heterozygous:

Gene, Chromosome & Base number	HGVS Genomic	HGVS Transcript	HGVS Protein	Number of homozygous LoF variants in ExAC	Reason for exclusion
NBPF1 1: 16902784	NC_000001.10:g.16902784G>T	NM_017940.4:c.2097C>A	NM_017940.4(*NBPF1*_i001):p.(Arg699Ser)	0	Does not pass truth sensitivity.
NBPF1 1: 16913608	NC_000001.10:g.16913608T>C	NM_017940.4:c.715A>G	NM_017940.4(*NBPF1*_i001):p.(Thr239Ala)	0	Does not pass truth sensitivity.
NBPF10 1: 145297661	NC_000001.10:g.145297661C>A	NM_001039703.3:c.536C>A	NM_001039703.3 (*NBPF10*_i001):p.(Ala179Asp)	38	Presence of homozygous LoF variants
NBPF10 1: 145299906	NC_000001.10:g.145299906T>G	NM_001039703.3:c.955T>G	NM_001039703.3 (*NBPF10*_i001):p.(Ser319Ala)	38	Presence of homozygous LoF variants
NBPF10 1: 145299925	NC_000001.10:g.145299925A>G	NM_001039703.3:c.974A>G	NM_001039703.3 (*NBPF10*_i001):p.(Gln325Arg)	38	Presence of homozygous LoF variants
FIGN 2: 164467319	NC_000002.11:g.164467320delT	NM_018086.2:c.1022delA	NM_018086.2(*FIGN*_i001):p.(Gln341Argfs*34)	0	Does not pass truth sensitivity.
FIGN 2: 164467322	NC_000002.11:g.164467322T>G	NM_018086.2:c.1020A>C	NM_018086.2(*FIGN*_i001):p.(Gln340His)	0	Does not pass truth sensitivity.
FIGN 2: 164467333	NC_000002.11:g.164467333T>C	NM_018086.2:c.1009A>G	NM_018086.2(*FIGN*_i001):p.(Ser337Gly)	0	Does not pass truth sensitivity.
FIGN 2: 164467336	NC_000002.11:g.164467336_164467337insTC	NM_018086.2:c.1005_1006insGA	NM_018086.2(*FIGN*_i001):p.(Tyr336Aspfs*40)	0	
FIGN 2: 164467339	NC_000002.11:g.164467340_164467342delTCC	NM_018086.2:c.1000_1002delGGA	NM_018086.2(*FIGN*_i001):p.(Gly334del)	0	
FIGN 2: 164467342	NC_000002.11:g.164467343_164467347insAAAAA	NM_018086.2:c.997_998insTTTTT	NM_018086.2(*FIGN*_i001):p.(Tyr333Phefs*44)	0	In a mononucleotide repeat.
CEP70 3: 138227578	NC_000003.11:g.138227579delT	NM_024491.2:c.945-193delA	(INTRONIC)	0	Intronic, no clear significance in causing disease

CEP70 3: 138229093	NC_000003.11:g.138229094_138229095insA	NM_024491.2:c.945-1709_945-1708insT	(INTRONIC)	0	Intronic, no clear significance in causing disease
CEP70 3: 138234309	NC_000003.11:g.138234310_138234311insA	NM_024491.2:c.945-6925_945-6924insT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
DUX4L4 4: 191003260	NC_000004.11:g.191003260G>T	NM_001177376.1:c.1171G>T	NM_001177376.1 (*DUX4L4*_i001):p.(Ala391Ser)	0	Does not pass truth sensitivity.
DUX4L4 4: 191003273	NC_000004.11:g.191003273A>T	NM_001177376.1:c.1184A>T	NM_001177376.1 (*DUX4L4*_i001):p.(Gln395Leu)	0	Does not pass truth sensitivity.
DUX4L4 4: 191003350	NC_000004.11:g.191003350C>A	NM_001177376.1:c.1261C>A	NM_001177376.1 (*DUX4L4*_i001):p.(Pro421Thr)	0	Does not pass truth sensitivity.
DUX4L4 4: 191003356	NC_000004.11:g.191003357_191003360insGAT	NM_001177376.1:c.1268_1269insGAT	NM_001177376.1 (*DUX4L4*_i001):p.(Pro423_Trp424insIle)	0	Does not pass truth sensitivity.
DUX4L4 4: 191003402	NC_000004.11:g.191003402C>A	NM_001177376.1:c.1313C>A	NM_001177376.1 (*DUX4L4*_i001):p.(Ala438Glu)	0	Does not pass truth sensitivity.
DAP 5: 10695977	NC_000005.9:g.10695978delT	NM_004394.2:c.153-12295delA	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
DAP 5: 10733308	NC_000005.9:g.10733308delG	NM_004394.2:c.152+14979delC	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
IQGAP2 5: 75709445	NC_000005.9:g.75709445_75709446insGG	NM_006633.2:c.46+10029_46+10030insGG	(INTRONIC)	0	Needs a second variant in the same gene to be plausible through this model.
IQGAP2 5: 75714050	NC_000005.9:g.75714050A>T	NM_006633.2:c.46+14634A>T	(INTRONIC)	0	Intronic, no clear significance in causing disease
IQGAP2 5: 75714418	NC_000005.9:g.75714419_75714421delGAA	NM_006633.2:c.46+15003_46+15005delGAA	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
IQGAP2 5: 75720513	NC_000005.9:g.75720513A>G	NM_006633.2:c.46+21097A>G	(INTRONIC)	0	Intronic, no clear significance in causing disease
IQGAP2 5: 75722730	NC_000005.9:g.75722731_75722743insGTGTGTGTGTGT	NM_006633.2:c.46+23315_46+23327insGTGTGTGTGTGT	(INTRONIC)	0	In a dinucleotide microsatellite repeat.
IQGAP2 5: 75733863	NC_000005.9:g.75733864delG	NM_006633.2:c.47-23531delG	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
IQGAP2 5: 75739294	NC_000005.9:g.75739295delT	NM_006633.2:c.47-18100delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
IQGAP2 5: 75756685	NC_000005.9:g.75756693delT	NM_006633.2:c.47-702delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.

IQGAP2 5: 75757192	NC_000005.9:g.75757194_75757195insA	NM_006633.2:c.47-201_47-200insA	(INTRONIC)	0	Intronic, no clear significance in causing disease
MRPS27 5: 71554088	NC_000005.9:g.71554088C>T	NM_015084.2:c.282-20133G>A	(INTRONIC)	0	Intronic, no clear significance in causing disease
MRPS27 5: 71566907	NC_000005.9:g.71566907C>G	NM_015084.2:c.281+24451G>C	(INTRONIC)	0	Intronic, no clear significance in causing disease
PKHD1 6: 51666454	NC_000006.11:g.51666454A>T	NM_138694.3:c.8303-10283T>A	(INTRONIC)	0	Intronic, no clear significance in causing disease
PKHD1 6: 51669268	NC_000006.11:g.51669268C>T	NM_138694.3:c.8303-13097G>A	(INTRONIC)	0	Intronic, no clear significance in causing disease
PKHD1 6: 51687189	NC_000006.11:g.51687189A>G	NM_138694.3:c.8302+8470T>C	(INTRONIC)	0	Intronic, no clear significance in causing disease
PRSS1 7: 142459689	NC_000007.13:g.142459689A>T	NM_002769.3:c.265A>T	NM_002769.3(*PRSS1*_i001):p.(Asn89Tyr)	0	Does not pass truth sensitivity.
PRSS1 7: 142459762	NC_000007.13:g.142459762T>A	NM_002769.3:c.338T>A	NM_002769.3(*PRSS1*_i001):p.(Leu113His)	0	Needs a second variant in the same gene to be plausible to be causing disease through this model.
PRSS1 7: 142459866	NC_000007.13:g.142459866G>C	NM_002769.3:c.442G>C	NM_002769.3(*PRSS1*_i001):p.(Ala148Pro)	0	Does not pass truth sensitivity.
PRSS1 7: 142459867	NC_000007.13:g.142459867C>T	NM_002769.3:c.443C>T	NM_002769.3(*PRSS1*_i001):p.(Ala148Val)	0	Does not pass truth sensitivity.
PRSS1 7: 142460779	NC_000007.13:g.142460779G>T	NM_002769.3:c.652G>T	NM_002769.3(*PRSS1*_i001):p.(Asp218Tyr)	0	Does not pass truth sensitivity.
PRSS1 7: 142460801	NC_000007.13:g.142460801A>G	NM_002769.3:c.674A>G	NM_002769.3(*PRSS1*_i001):p.(Lys225Arg)	0	Does not pass truth sensitivity.
TRBV10-1 7: 142231861	NC_000007.13:g.142231861C>T	No Transcripts found in variant region		9254	T-cell receptors are normally very variable.
TRBV10-1 7: 142231977	NC_000007.13:g.142231977C>T	No Transcripts found in variant region		9254	T-cell receptors are normally very variable.
ANKRD20A1 9: 67934829	NC_000009.11:g.67934829G>A	NM_032250.3:c.599G>A	NM_032250.3(*ANKRD20A1*_i001):p.(Arg200Gln)	0	Does not pass truth sensitivity.
ANKRD20A1 9: 67968657	NC_000009.11:g.67968657G>A	NM_032250.3:c.2216G>A	NM_032250.3(*ANKRD20A1*_i001):p.(Arg739His)	0	Both parents also appear to have variant, does not pass truth sensitivity.
ANKRD20A1 9: 67968720	NC_000009.11:g.67968720T>C	NM_032250.3:c.2279T>C	NM_032250.3(*ANKRD20A1*_i001):p.(Ile760Thr)	0	Does not pass truth sensitivity.

ANKRD20A4 9: 69423668	NC_000009.11:g.69423668A>C	NM_001098805.1:c.1964A>C	NM_001098805.1 (*ANKRD20A4*_i001):p.(Asp655Ala)	4	Does not pass truth sensitivity. Presence of homozygous LoF variants
ANKRD20A4 9: 69423920	NC_000009.11:g.69423920G>A	NM_001098805.1:c.2216G>A	NM_001098805.1 (*ANKRD20A4*_i001):p.(Arg739His)	4	Presence of homozygous LoF variants
PIK3AP1 10: 98392164	NC_000010.10:g.98392165_98392167delTTT	NM_152309.2:c.1376-3917_1376-3915delAAA	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
PIK3AP1 10: 98397075	NC_000010.10:g.98397075delA	NM_152309.2:c.1375+8155delT	(INTRONIC)	0	Deletion overlapped by a common deletion.
PIK3AP1 10: 98400843	NC_000010.10:g.98400855_98400857insGT	NM_152309.2:c.1375+4373_1375+4375insAC	(INTRONIC)	0	In a dinucleotide microsatellite repeat.
CDK4 12: 58143936	NC_000012.11:g.58143937_58143949delAATAAATAAATA	NM_000075.2:c.632+490_632+502delTATTTATTTATT	(INTRONIC)	0	Two variants in the same gene necessary to be causal in this model.
CDK4 12: 58143979	NC_000012.11:g.58143980_58143983delAAAA	NM_000075.2:c.632+456_632+459delTTTT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
CDK4 12: 58143983	NC_000012.11:g.58143983delA	NM_000075.2:c.632+456delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
POTEM 14: 20014669	NC_000014.8:g.20014669C>T	NM_001145442.1:c.640G>A	NM_001145442.1 (*POTEM*_i001):p.(Val214Ile)	0	Does not pass truth sensitivity.
POTEM 14: 20019757	NC_000014.8:g.20019757T>C	NM_001145442.1:c.464A>G	NM_001145442.1 (*POTEM*_i001):p.(Lys155Arg)	0	Does not pass truth sensitivity.
GOLGA8N 15: 32891501	NC_000015.9:g.32891501C>T	XM_003959944.2:c.665C>T	XM_003959944.2 (*GOLGA8R*_i001):p.(Ala222Val)	0	Many reads with a mapping quality of 0.
GOLGA8N 15: 32895892	NC_000015.9:g.32895892A>T	XM_003959944.2:c.1697A>T	XM_003959944.2 (*GOLGA8R*_i001):p.(Glu566Val)	0	Does not pass truth sensitivity.
GOLGA8N 15: 32895903	NC_000015.9:g.32895903G>A	XM_003959944.2:c.1708G>A	XM_003959944.2 (*GOLGA8R*_i001):p.(Ala570Thr)	0	Does not pass truth sensitivity.
GOLGA8Q 15: 30848826	NC_000015.9:g.30848826C>T	XM_001126407.1:c.514C>T	XM_001126407.1 (*LOC727909*_i001):p.(His172Tyr)	1	Does not pass truth sensitivity.
GOLGA8Q 15: 30854100	NC_000015.9:g.30854100T>C	XM_001126407.1:c.1475T>C	XM_001126407.1 (*LOC727909*_i001):p.(Ile492Thr)	1	Does not pass truth sensitivity.
PARN 16: 14700638	NC_000016.9:g.14700639delA	NM_001134477.2:c.477-256delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
PARN 16: 14555222	NC_000016.9:g.14555222A>C	NM_001134477.2:c.1488-14284T>G	(INTRONIC)	0	Two variants in the same gene necessary to be causal in this model.
PKD1 16: 2143546	NC_000016.9:g.2143546C>T	NM_000296.3:c.11012G>A	NM_000296.3(*PKD1*_i001):p.(Arg3671Gln)	0	
PKD1 16: 2159667	NC_000016.9:g.2159667T>C	NM_000296.3:c.5501A>G	NM_000296.3(*PKD1*_i001):p.(Asn1834Ser)	0	

CDK12 17: 37631800	NC_000017.10:g.37631801_37631802insT	NM_015083.1:c.1931+3785_1931+3786insT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
CDK12 17: 37636517	NC_000017.10:g.37636518_37636519insT	NM_015083.1:c.1931+8502_1931+8503insT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
CDK12 17: 37637880	NC_000017.10:g.37637881_37637882delCG	NM_015083.1:c.1932-8929_1932-8928delCG	(INTRONIC)	0	Needs a second variant in the same gene to be plausible through this model.
CDK12 17: 37641422	NC_000017.10:g.37641423_37641424insAA	NM_015083.1:c.1932-5387_1932-5386insAA	(INTRONIC)	0	In a dinucleotide microsatellite repeat.
ANKLE1 19: 17397499	NC_000019.9:g.17397500_17397501delTT	NM_001278444.1:c.1933_1934delTT	NM_001278444.1 (*ANKLE1*_i001):p.(Leu645Valfs*33)	17	In a mononucleotide repeat.
ANKLE1 19: 17397501	NC_000019.9:g.17397501_17397503delinsGTG	NM_001278444.1:c.1933_1935delinsGTG	NM_001278444.1 (*ANKLE1*_i001):p.(Leu645Val)	17	In a dinucleotide microsatellite repeat.
FRG1B 20: 29631613	NC_000020.10:g.29631613T>C	NR_003579.1:n.709T>C	NR_003579.(*FRG1B*_001):p.(Cys237Arg)	0	Does not pass truth sensitivity.
FRG1B 20: 29633902	NC_000020.10:g.29633902C>A	NR_003579.1:n.841C>A	NR_003579.(*FRG1B*_001):p.(Pro281Thr)	0	Does not pass truth sensitivity.
SLCO4A1 20: 61291767	NC_000020.10:g.61291768_61291769delGA	NM_016354.3:c.893_894delAG	NM_016354.3(*SLCO4A1*_i001):p.(Glu298Alafs*254)	0	Invalid after sequencing by the Sanger methodology.
SLCO4A1 20: 61291771	NC_000020.10:g.61291772delT	NM_016354.3:c.896delT	NM_016354.3(*SLCO4A1*_i001):p.(Leu299Argfs*46)	0	Invalid after sequencing by the Sanger methodology.

Homozygous:

Gene, Chromosome & Base number	HGVS Genomic	HGVS Transcript	HGVS Protein	Number of homozygous LoF variants in ExAC	Reason for exclusion
ATN1 12: 7045891	NC_000012.11:g.7045898_7045911delCAGCAGCAGCAGCAG	NM_001007026.1:c.1468_1481delCAGCAGCAGCAGCAG	NM_001007026.1 (*ATN1_i001*):p.(Gln498_Gln502del)	164	Inherited repeat at a polymorphic locus, length in the normal range
AP001055.1 21: 45588080	NC_000021.8:g.45588080_45588115delGAGAAGCTGTAGGATCCACACCGCCTTTCCGTACAC	XM_001722818.1:c.755_790delGTGTACGGAAAGGCGGTGTGGATCCTACAGCTTCTC	XM_001722818.1 (*LOC100129890*_i001):p.(Arg252_Ser263del)	NA	XM is a hypothetical transcript. No direct evidence for its existence.
FAM231B 1: 16865770	NC_000001.10:g.16865771delT	XM_001721533.2:c.214delT	XM_001721533.2 (*FAM231B*_i001):p.(Trp72Glyfs*29)	NA	XM is a hypothetical transcript. No direct evidence for its existence.
NUMBL 19: 41173895	NC_000019.9:g.41173895_41173898delCTGT	NM_004756.3:c.1305_1308delACAG	NM_004756.3(*NUMBL*_i001):p.(Gln435Hisfs*149)	0	An upstream deletion containing this deletion has a frequency of 0.6447, too common to be causative.

CABP1 12: 121093631	NC_000012.11:g.121093631_121093633delTGC	NM_001033677.1:c.655-4050_655-4048delTGC	(INTRONIC)	55	An upstream deletion containing this deletion has a frequency of 0.4773, too common to be causative.
LY75 2: 160687099	NC_000002.11:g.160687099_160687100insT	NM_002349.3:c.3958+1081_3958+1082insA	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
MRPS27 5: 71538492	NC_000005.9:g.71538499_71538502delACAC	NM_015084.2:c.282-4548_282-4545delGTGT	(INTRONIC)	0	In a dinucleotide microsatellite repeat.
MRPS27 5: 71540697	NC_000005.9:g.71540697_71540698delCT	NM_015084.2:c.282-6744_282-6743delAG	(INTRONIC)	0	In a dinucleotide microsatellite repeat.
MRPS27 5: 71566020	NC_000005.9:g.71566021_71566026delAATAAT	NM_015084.2:c.281+25332_281+25337delATTATT	(INTRONIC)	0	Inherited polymorphic trinucleotide repeat
MRPS27 5: 71571858	NC_000005.9:g.71571858G>C	NM_015084.2:c.281+19500C>G	(INTRONIC)	0	Intronic, no clear significance in causing disease
MRPS27 5: 71582401	NC_000005.9:g.71582403_71582404insA	NM_015084.2:c.281+8954_281+8955insT	(INTRONIC)	0	Intronic, no clear significance in causing disease
MRPS27 5: 71584229	NC_000005.9:g.71584229G>A	NM_015084.2:c.281+7129C>T	(INTRONIC)	0	Intronic, no clear significance in causing disease
IQGAP2 5: 75724901	NC_000005.9:g.75724903_75724904insAA	NM_006633.2:c.46+25487_46+25488insAA	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
IQGAP2 5: 75733844	NC_000005.9:g.75733844T>G	NM_006633.2:c.47-23551T>G	(INTRONIC)	0	Intronic, no clear significance in causing disease
AC073333.1 7: 16829092	NC_000007.13:g.16829092A>G	NM_006408.3:c.*3440T>C		NA	Does not pass truth sensitivity.
NOP9 14: 24769849	NC_000014.8:g.24769850_24769851insGAG	NM_174913.1:c.484_485insGAG	NM_174913.1(*NOP9*_i001):p.(Ala161_Glu162insGly)	232	Presence of homozygous LoF variants
SPATA31E3P 15: 23470965	NC_000015.9:g.23470965C>T	No Transcripts found		NA	Excluded pseudogenes.
PARN 16: 14551654	NC_000016.9:g.14551655delA	NM_001134477.2:c.1488-10717delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
SSH2 17: 28084745	NC_000017.10:g.28084747delA	NM_033389.2:c.107+36165delT	(INTRONIC)	0	In a mononucleotide microsatellite repeat.
PRKACA 19: 14209300	NC_000019.9:g.14209300A>T	NM_002730.3:c.420-598T>A	(INTRONIC)	0	In a dinucleotide microsatellite repeat