2 Exome sequencing improves genetic diagnosis of structural fetal abnormalities

2.1 Introduction

2.1.1 The impact and causes of fetal structural abnormalities

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

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

2.1.2 Current techniques for prenatal genetic diagnosis

Sampling methods

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

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

Karyotyping

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

An alternative to classical cytogenetics is molecular cytogenetics, such as fluorescence *in situ* hybridisation (FISH). During this technique, fluorescent-tagged oligonucleotide probes complementary to a DNA sequence of interest are used to visualise whole chromosomes. It was first developed in the 1980s (28), and subsequent developments include chromosome ‘paints’ based on unique, chromosome-specific sequences which

allow each chromosome to be visualised simultaneously in a different colour (29). Known as spectral karyotyping, this has some advantages over Giemsa banding in that it allows easy identification of the chromosomal origin of genetic material, and it has a higher resolution of 1-2 Mb (30). However, it is usually used in conjunction with other methods, as it has the major disadvantage of not being able to detect intrachromosomal aberrations.

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

Microarrays and quantitative fluorescent PCR

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

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

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

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

Non-invasive prenatal testing

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

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

Two proof of concept studies published in 2012 showed that it is possible to sequence the whole genome of a fetus non-invasively using cfDNA, to a sufficient depth to be

able to call inherited SNVs, using parental haplotypes to distinguish fetal from maternal variants (53, 54). However, the sensitivity and specificity of the SNV calling are as yet insufficient to consider using this approach in clinical practice.

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

2.1.3 Next generation sequencing

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

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

NGS is increasingly being used in the clinical setting, as a diagnostic test for patients with rare diseases. Often, exome sequencing is used. However, the most appropriate method depends upon the phenotype. For example, retinal dystrophy (RD) is a rare,

inherited, degenerative cause of visual impairment and blindness. It is genetically heterogeneous, but a higher proportion of RD-associated genes have been identified, than for other phenotypes. Sequencing of 105 RD-associated genes therefore has a diagnostic yield of 55% (60). In contrast, exome sequencing of patients with rare, undiagnosed, developmental diseases typically has a diagnostic yield of around 25% (11, 61). Therefore, for phenotypes like RD, NGS using gene panels might be a more cost-efficient diagnostic method than exome sequencing.

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

2.1.4 Variant prioritisation strategies

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

The next step depends on whether the application of the sequencing is clinical diagnostics, or research. For clinical diagnostics, matches between a gene that contains a variant in the patient, and genes that are known to be associated with the phenotype of that patient, are identified. For research, novel disease-associated genes

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

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

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

2.1.5 Prenatal next generation sequencing: proof of concept

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

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

2.1.6 Aims, context, and colleagues

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

The overall aims of this project were to use exome sequencing on a cohort of fetuses with structural abnormalities, and their parents, to estimate the diagnostic yield of this technique for this purpose, and to identify any issues that would need to be addressed

prior to exome sequencing being implemented as a gene discovery or diagnostic tool for structural fetal abnormalities on a large scale.

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

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

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

2.2 Methods

2.2.1 Cohort

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

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

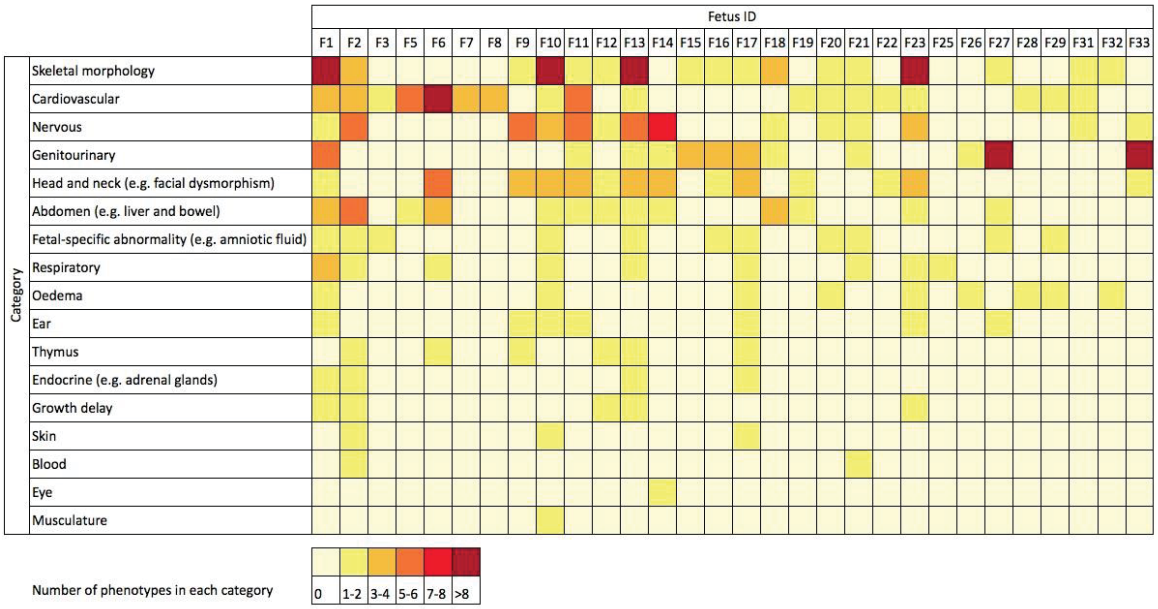

Figure 2 - 1: Matrix showing categories of phenotypes in the cohort of fetuses with structural abnormalities
For each fetus (F1-F33), the colour indicates the number of observed phenotypes that are in each category of phenotypes. For example, F1 has more than eight separate abnormalities of skeletal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

2.2.2 Exome sequencing

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

2.2.3 VCF file merging, annotation, and quality control

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

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

2.2.4 Identification of *de novo* SNVs and indels

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

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

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

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

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

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

2.2.6 Identification of CNVs

CoNVex detects copy number variation from exome data using comparative read depth. (ftp://ftp.sanger.ac.uk/pub/users/pv1/CoNVex/) It corrects for technical variation between samples and detects copy number variable segments using a heuristic error-weighted score and the Smith-Waterman algorithm. It detects deletions and

duplications of targeted sequences from few hundred base pairs in size to a few megabases or more.

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

2.2.7 Sanger sequencing

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

2.2.8 Interpretation of variants

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

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

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

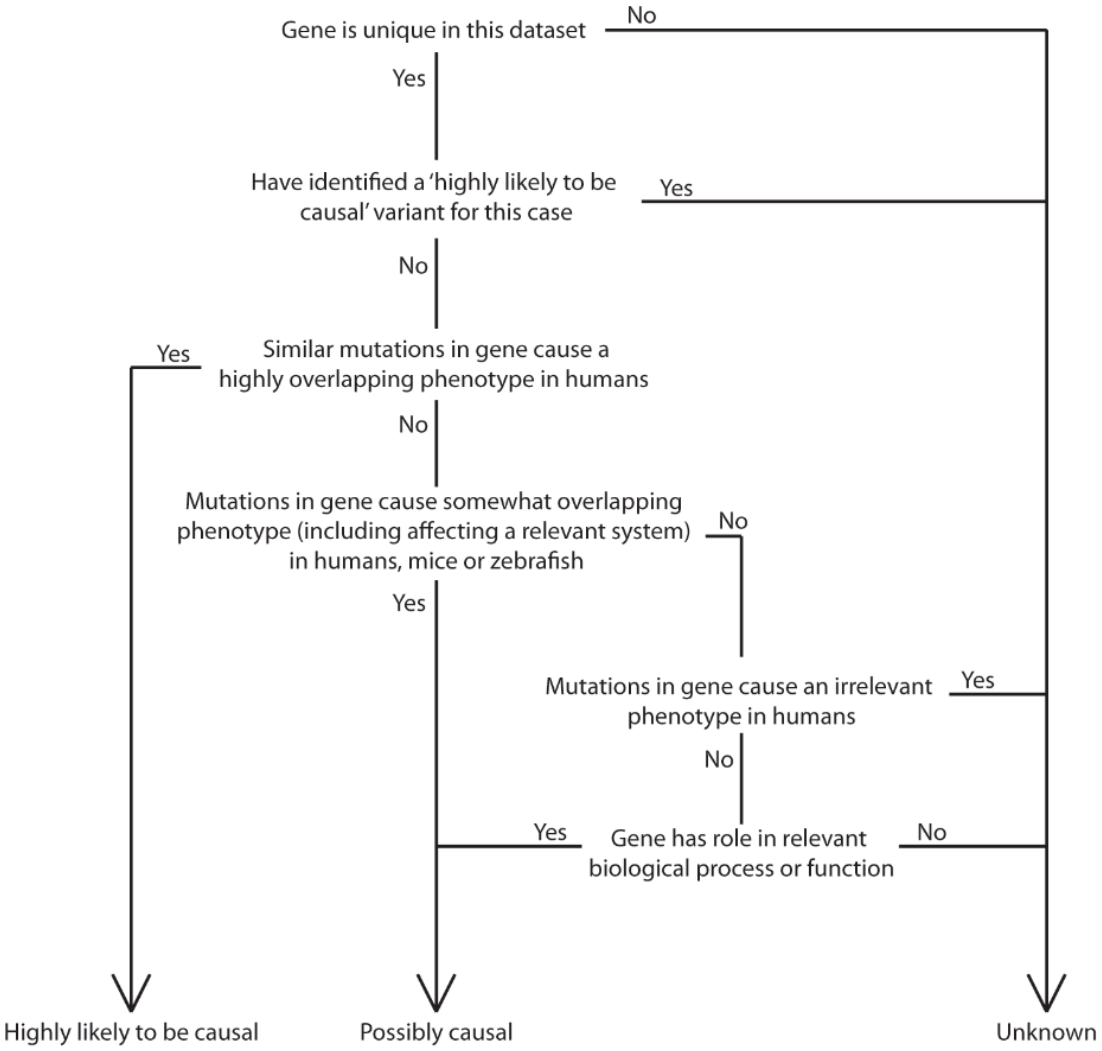

Figure 2 - 2: Decision tree for classifying candidate genes into three categories.
Data were used where available from the following sources: Online Mendelian Inheritance in Man (OMIM), DDG2P, Biology Gene Portal System (BioGPS), National Human Genome Research Institute (NHGRI) genome-wide association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

2.3 Results

2.3.1 The exome sequencing data are of high quality

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

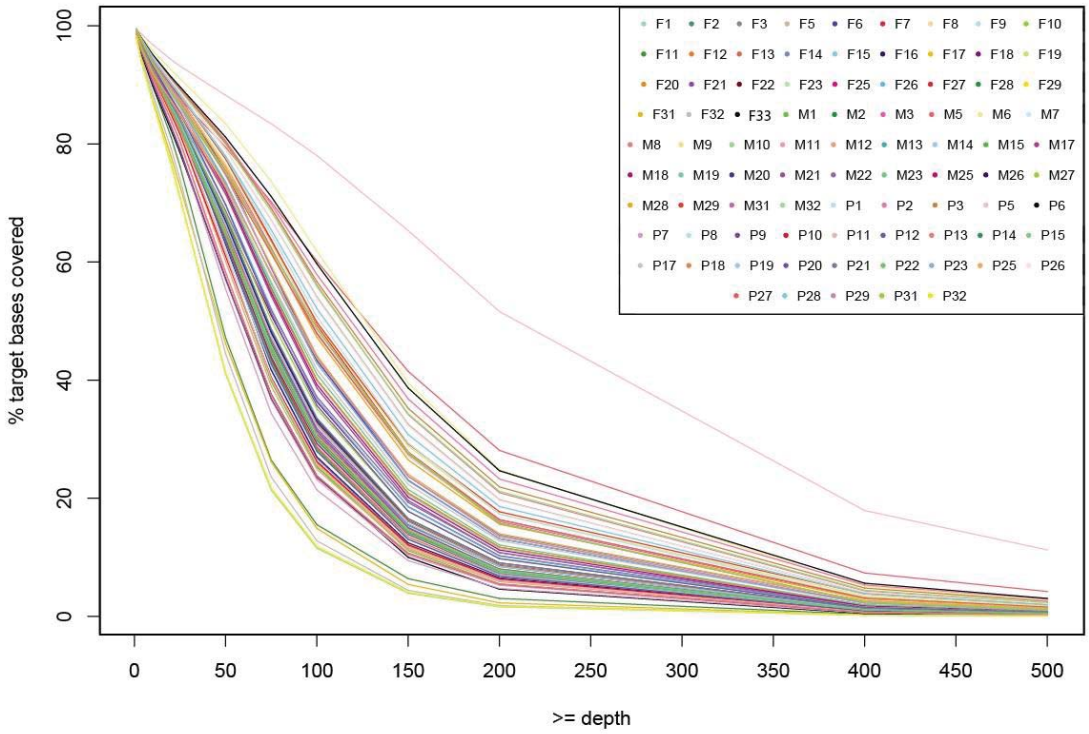

Figure 2 - 3: Target coverage of exome sequencing reads by sample.
P5 has higher coverage, as it was not sequenced as part of a pool. This figure and legend have been published (73).

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

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

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

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

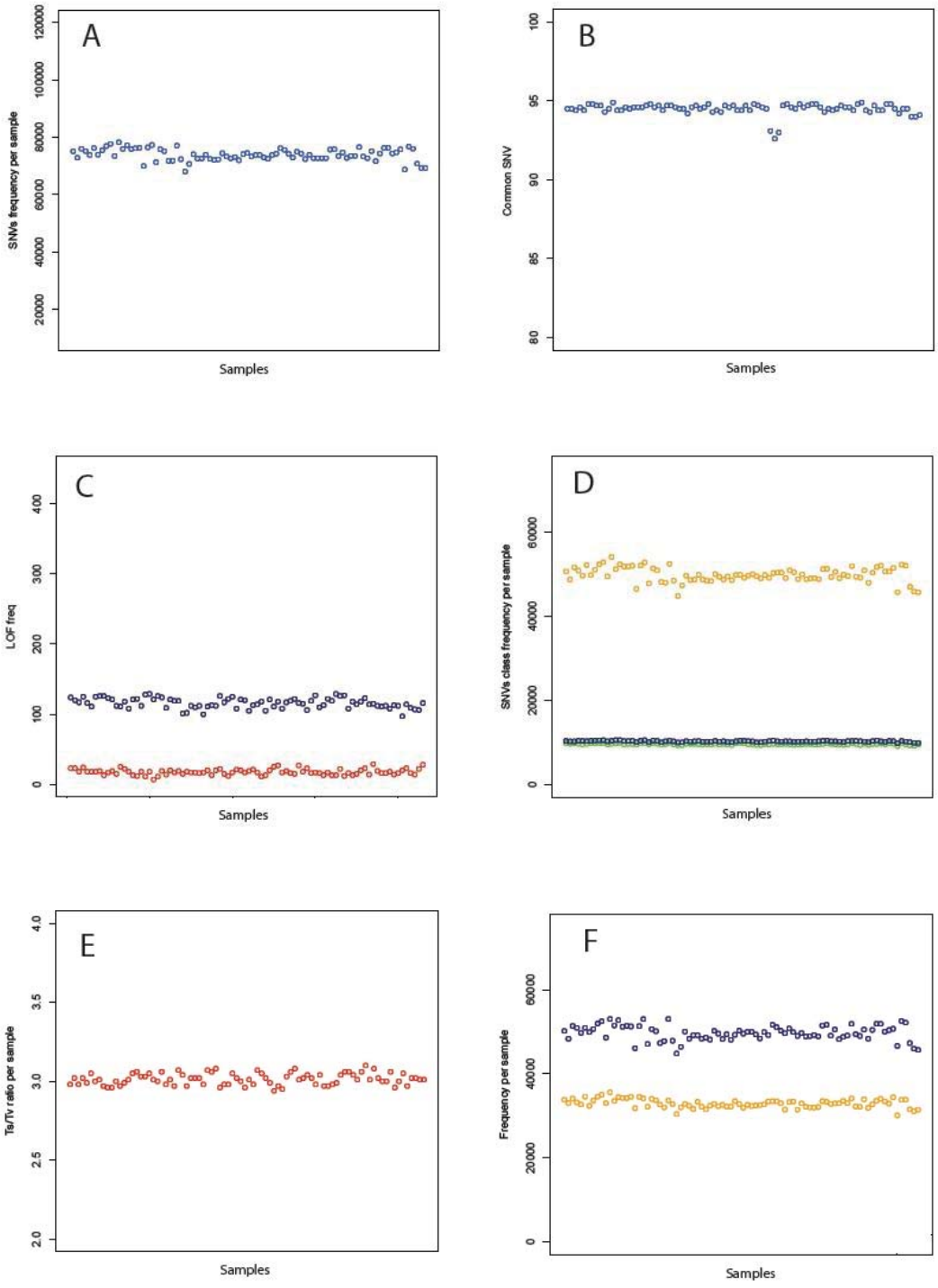

Figure 2 - 4: Quality control metrics for single nucleotide variants.
(A) Number of high-quality SNVs per sample. **(B)** Percent of SNVs that are common (≥1% population frequency) per sample. The cluster of three samples with a lower percentage of common SNVs represents F19, M19 and P19. These individuals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

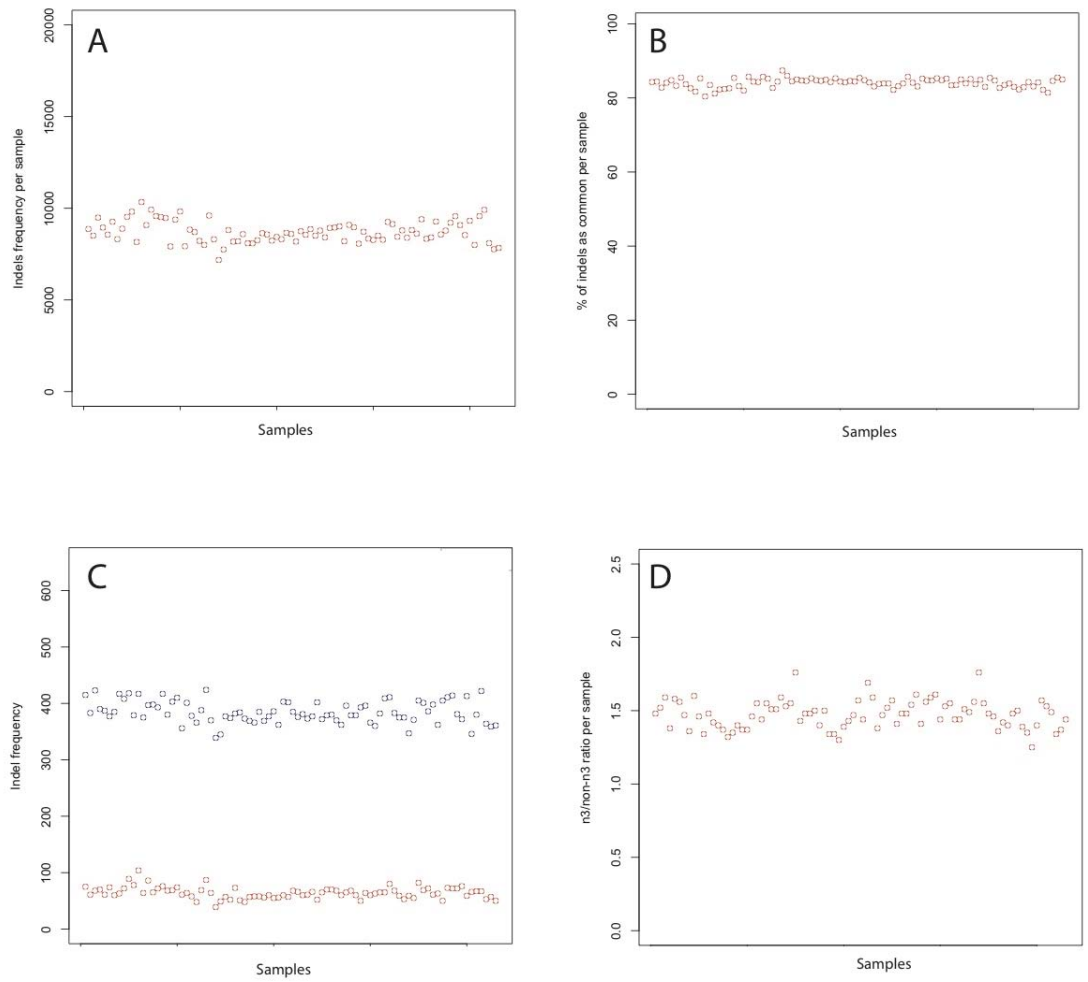

Figure 2 - 5: Quality control metrics for indels.
(A) Number of high-quality indels per sample. **(B)** Percent of indels that are common (≥1% population frequency) per sample. **(C)** Number of coding indels per sample. Common (≥1%) are shown in blue and rare (<1%) are shown in red. **(D)** Ratio of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

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

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

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

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

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

The expected percentage of *de novo* mutations in coding or splicing sequence that are synonymous is 29% (83), however, I observed that only three (9%) of the 33 validated independent *de novo* mutations were synonymous, with 26 being non-synonymous, three nonsense and one in a splice site. Thus the proportion of validated *de novo* mutations that are predicted to have a functional consequence of the encoded protein

is significantly enriched over what would be expected by chance (p=0.007), suggesting that an appreciable subset of these functional mutations is likely to be pathogenic. For two of the *de novo* mutations, the proportion of reads that support the alternative allele was significantly less than the expected 50% for a non-mosaic, heterozygous mutation. This provides suggestive evidence that these mutations are mosaic. These mutations were c.313G>C (p.105E>Q) in *PARD3B* (ENST00000349953) in F9, and c.2921G>T (p. 974C>F) in *SEC31B* (MIM 610258, ENST00000370345) in F33 (Table 2-2).

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

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

ID	CHR	Start position	End position	Size (kb)	CNV type	Inheritance model	Gene
F14	X	13770686	13791294	20.6	DEL	*de novo*	*GPM6B; OFD1*
F19	X	48155306	48270940	115.6	DUP	Inherited X linked	*SSX3; SSX4; SSX4B*
F3	X	103267111	103301913	34.8	DUP	Inherited X linked	*H2BFM; H2BFWT*

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

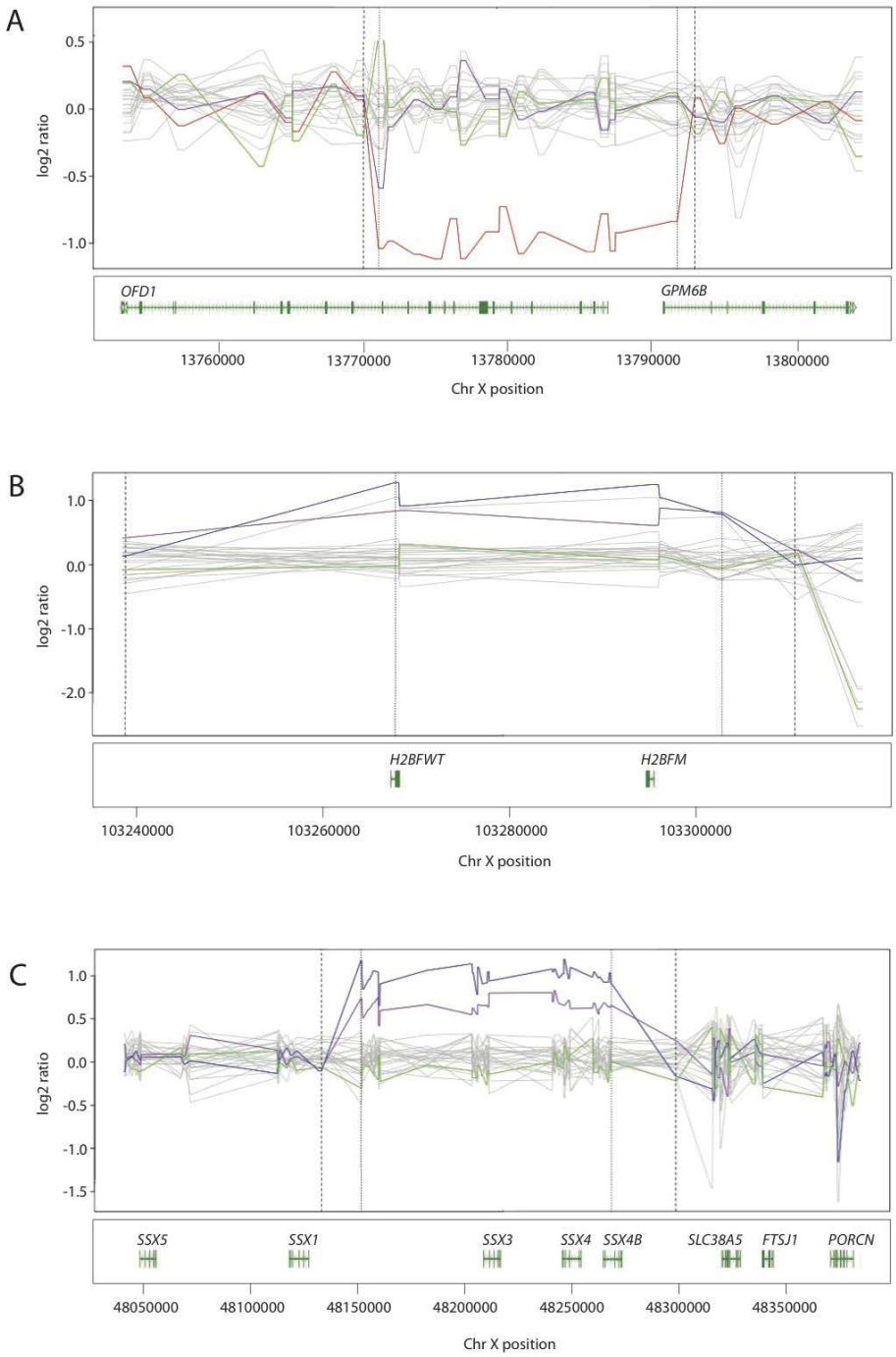

Figure 2 - 6: Log_2 ratios of candidate CNVs in fetuses with structural abnormalities.
(A) F14; **(B)** F19; **(C)** F3. In each plot the x-axis indicates the genomic coordinates. The top panel indicates the normalised log_2 ratio of the exome read depth, compared to a group of controls. The red line shows the log_2 ratio of the fetus, where the variant is a deletion, and the blue line shows the log_2 ratio of the fetus where the variant is a duplication. The purple line shows the log_2 ratio of the mother, and the green line shows the log_2 ratio of the father. The grey lines show the log_2 ratio of control samples. The vertical small dashed lines show the minimum deleted/duplicated region and the vertical wide dashed lines show the maximum deleted/duplicated region. The bottom panel shows the protein-coding genes present in each region. This figure and legend have been published (73).

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

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

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

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

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

Two of the *de novo* SNVs are highly likely to be pathogenic, and two are possibly causal. One *de novo* mutation that is highly likely to be causal was found in F23, a male fetus with features consistent with thanatophoric dysplasia, including a large head, disproportionately short limbs, and a narrow, bell-shaped chest. I found the

missense mutation c.1118A>G (p.373Y>C) in fibroblast growth factor receptor 3 (*FGFR3*, MIM 134934 (http://www.omim.org/), ENST00000440486) (Figure 2-7). FGFR3 is a well-characterised negative regulator of bone growth, missense mutations in which are known to cause a wide range of skeletal dysplasias, most commonly achondroplasia. There is a very tight correlation between specific *FGFR3* mutations, and the phenotype, for a review see (91). The mutation p.373Y>C is known to cause thanatophoric dysplasia (23), giving high confidence that c.1118A>G in *FGFR3* is the causative mutation in F23.

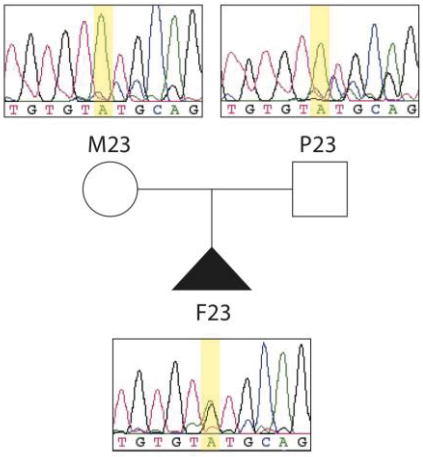

Figure 2 - 7: Pedigree of trio 23, showing Sanger sequencing of *de novo* mutation in *FGFR3*.

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

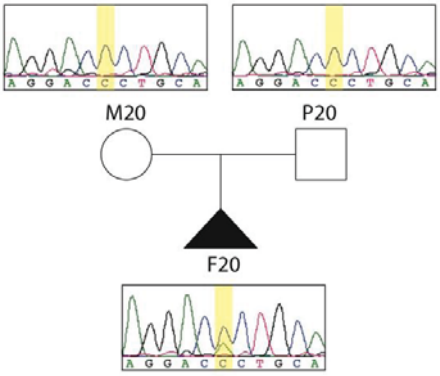

Figure 2 - 8: Pedigree of trio 20, showing Sanger sequencing of *de novo* mutation in *COL2A1*.

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

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

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

mutation, despite appearing possibly clinically relevant, is unlikely to be significantly pathogenic.

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

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

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

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

from knockdown of *PARD3* and the phenotypes in F9 and F19 is interesting, however I judged that the current knowledge of the function of *PARD3B* is insufficient to categorise the mutations identified in our cohort as being possibly causal.

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

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

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

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

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

In F6, whose laterality phenotype has been described, I found the compound heterozygous variants c.4264G>A (p.1422V>M) and c.3686G>A (p.1229R>Q) in *RERE* (MIM 605226, ENST00000337907). *RERE*, which is in the retinoic acid pathway,

has a role in establishing bilateral symmetry. Although it is not a known human disease-associated gene, homozygous knockout mice develop asymmetrically and have cardiovascular outflow defects. Homozygous zebrafish mutants have cartilage and skeletal defects, abnormal fins and otoliths, reduced viability, deformed brains, and absent gills (107-109). In total I have identified two genes with *de novo* mutations and one gene with inherited variants that could possibly account for the phenotype in F6. It is not possible to say which is most likely to be causative, as none of the candidate genes are known to harbour variants that cause the exact phenotype reported here. One possibility is that multiple variants contribute to this multisystemic phenotype, as has been reported in other exome sequencing studies of rare disease (3, 11).

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

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

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

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

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

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

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

score. For this experiment, all candidate genes with a maximum eXtasy score >0.5 also have a combined p-value of $< 5.6 \times 10^{-6}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

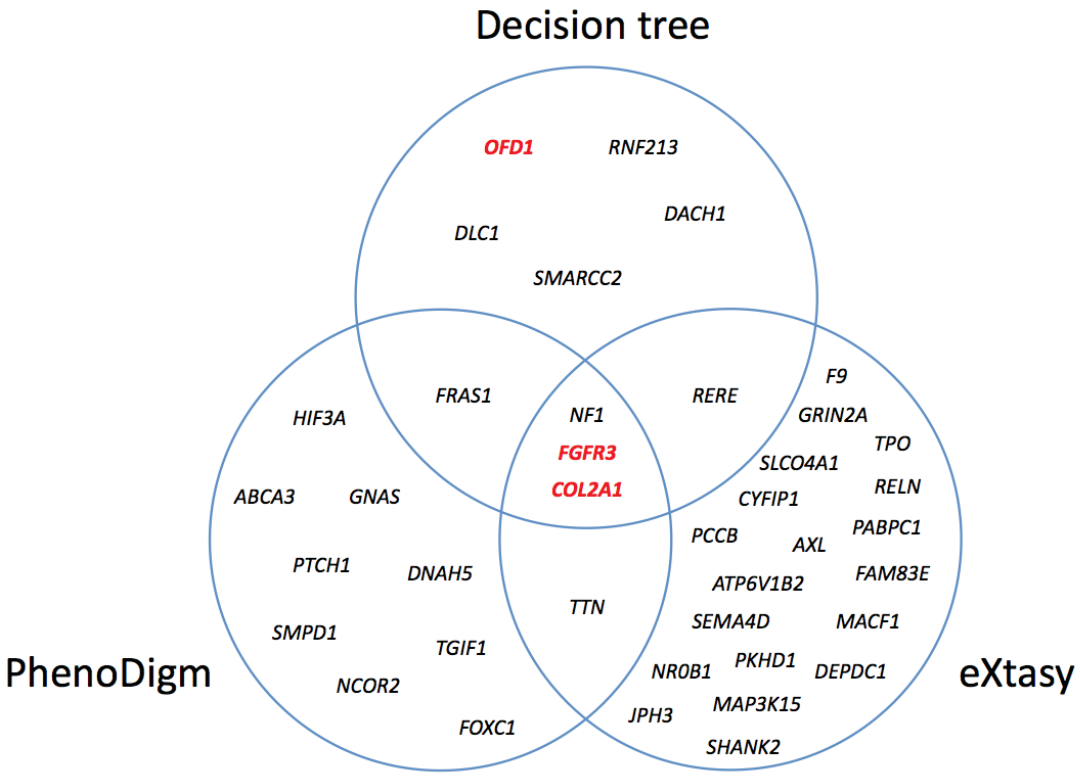

Figure 2 - 9: Venn diagram showing overlap between the genes prioritised by each of the three methods.
The genes named in the decision tree circle include both 'highly likely to be causal' and 'possibly causal' candidates, with the former in red.

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

2.3.13 The continuing need for manual curation

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

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

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

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

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

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

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

F19 has a high number of inherited, apparently rare variants (Appendix 3). F19 is of Indian ancestry, whereas the majority of the cohort is of European ancestry. It is likely therefore that some of the apparently rare variants that I have identified in F19 are in

fact more common in this population, but I have not been able to identify them as such due to an underrepresentation of individuals of Indian ancestry in the databases I used to filter the variants.

2.3.15 The estimated diagnostic yield of this study is 10%

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

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

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

Table 2 - 6: Summary of all candidate genes identified in 30 fetuses with structural abnormalities.
Column headers indicate the type of variant associated with the candidate genes. Bold red text indicates variants that are highly likely to be causal. Bold orange text indicates variants that are possibly causal. Square brackets contain genes in a single CNV. [1]Monozygotic twins; [2]Siblings; [3]Synonymous *de novo* mutation; [4]Indian ancestry.

2.4 Discussion

2.4.1 Summary

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

2.4.2 The diagnostic yield in context

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

Nevertheless, our diagnostic rate is lower than that found in exome sequencing studies of rare postnatal diseases, which is typically around 25% (3, 11, 61). There are several possible reasons for this. First, our estimate of 10%, being based on a relatively small sample size, has a broad confidence interval of 3.5% - 25.6%, meaning that a diagnostic rate of up to 25% could be possible in prenatal samples, and the diagnostic rate in this study might just be lower just by chance. Second, it is likely that in some cases, variants in the same gene will have different phenotypic manifestations between prenatal and postnatal stages of development (124). It seems likely for example, that, for a given variant or gene, one might observe more severe phenotypes *in utero*, which may not be compatible with life postnatally. Given that I interpreted the data in this

study by comparing fetal phenotypes to available data, the vast majority of which is postnatal, this makes interpretation more difficult. Similarly, for some of the fetuses in this study the only phenotypic data came from ultrasound scans. There are many phenotypes that cannot be identified from an ultrasound scan including subtle morphological abnormalities, most metabolic phenotypes, and behavioural and cognitive deficits. This potentially incomplete phenotype data also complicates variant interpretation.

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

2.4.3 Comparison of variant interpretation methods

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

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

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

Of the ten variants that I initially classified as highly likely to be causal or possibly causal, PhenoDigm also highlighted four of them as interesting. This is a promising degree of overlap. The main disadvantage of PhenoDigm is that if there is no animal model for a particular gene, it cannot be used. This limits its utility in practice, and means that it could not be used as the sole method of variant prioritisation, at least until a higher proportion of mouse genes have phenotyped knockouts. Similarly, there are cases where the phenotype of a human and the phenotype of a mouse with a variant on the orthologue of the same gene are not similar (125). While these cases are not

typical, they could lead to misleading results from PhenoDigm. The other disadvantage of PhenoDigm is that it can give a very significant score for a gene when there is overlap of a single phenotype. But this does not equate to a high degree of overall phenotypic overlap. For example, PhenoDigm identified *ABCA3* as an interesting candidate in F17, because the mouse has a similar lung defect to the fetus. However, the fetus had 14 phenotypes, only two of which overlapped with the mouse. Almost all of the phenotypes of the mouse model had to do with the lungs, whereas the fetus had many additional affected systems that did not recapitulate in the mouse model. Therefore, I concluded that the *ABCA3* variants are unlikely to be causal.

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

2.4.4 The ethics of next generation sequencing for prenatal genetic diagnosis

The many thorny ethical issues surrounding NGS in the clinical context have been extensively debated, chief among them are whether to report incidental findings, and how to report variants of unknown significance (VOUS) (126). In the prenatal context, the issues are similar but amplified, partly due to the possibility of termination of the

pregnancy. One possible application of prenatal sequencing that raises some unique ethical questions is widespread use in the general population.

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

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

Another question is whether return to families information on variants that are likely to cause late-onset disease, or have incomplete penetrance, such as a *BRCA1* variant

that confers an 80% risk of developing cancer later in life (129). Some argue that families have a right to this information to do with what they will, even if it will result in increased termination rates, and termination of some healthy fetuses (130). An alternative is to do more targeted sequencing based on the indication for the test, so as to avoid incidental findings.

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

2.4.5 Next generation sequencing is the future of prenatal genetic diagnostics

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

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

Exome sequencing is currently considered more cost-efficient than whole-genome sequencing for clinical diagnostic purposes. However, for several reasons, I predict an

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

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

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

In conclusion, the main outcomes of this project are as follows. We have achieved an approximate diagnostic yield of 10% in this small cohort. All of these 10% were *de*

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