3 | Genetic investigations of Tetralogy of Fallot in trios

Collaboration note

Dr. Sebastian Gerety and Dr. Sarah Lindsay generated some of the data described in this chapter. Sebastian performed the gene knockdown in zebrafish (appendix A) while Sarah provided technical assistance for the validation experiments for de novo mutations using PCR and capillary sequencing.

3.1 Introduction

3.1.1 Historical overview on Tetralogy of Fallot

In 1888, Étienne-Louis Arthur Fallot, a French physician, described heart anatomical features and linked them to the clinical presentation of a "*la maladie bleue*" or "the blue disease" [293]. Fallot noticed an interventricular communication, sub pulmonary stenosis, biventricular origin of the aorta and hypertrophy of the right ventricle in three patients with cyanotic discoloration. Today, we are aware that others such as Stenonis (1672), Farre (1814), Peacock (1866) and von Rokitansky (1875) also observed these anatomical features prior to Arthur Fallot. However, Fallot was the first to correlate these findings to the clinical features [294]. In 1924, Maude Abbott coined the term "Tetralogy of Fallot" (ToF) as a convenient for of identification instead of listing all four anatomical features [295] in her "Atlas of Congenital Cardiac Disease" [2, 296].

3.1.2 Epidemiology and recurrence risks of Tetralogy of Fallot

Tetralogy of Fallot occurs in 3 out of every 10,000 live births, and accounts for 10% of all CHD cases and is considered to be one of the most common cyanotic cardiac lesions beyond neonatal age [297]. Both genders are equally affected [298], but a recent report from the PAN study, a nation-wide study in Germany, showed that slightly more males are affected than females (1.4:1) [64]. A few

risk factors have been identified that increase the risk of ToF such as the age of father ≥ 25 [299], race and ethnicity may also contribute to differences in the prevalence of ToF. Compared to black infants, white infants were found to have an increased prevalence of many CHD subtypes including ToF [300].

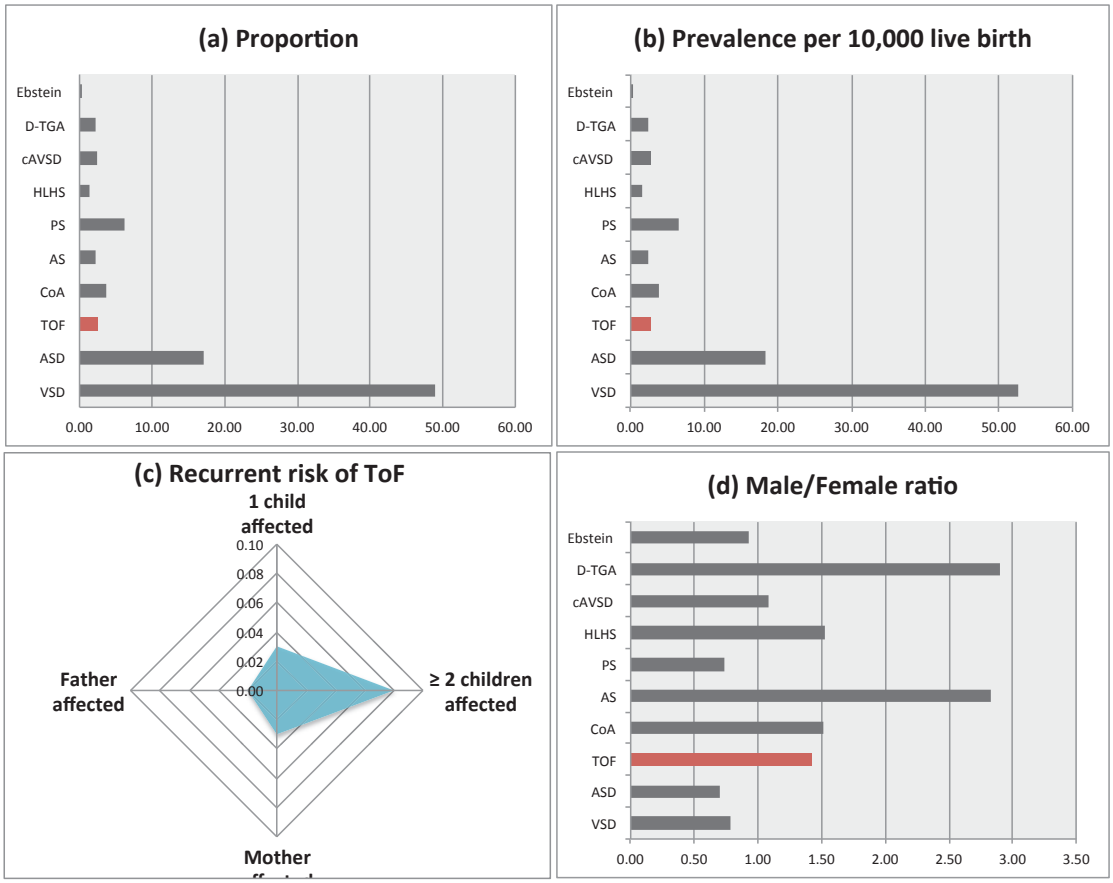


Figure 3-1 (a) Proportion of different CHD subtypes, including Tetralogy of Fallot (red bar) in the PAN registery (n=7,245) during one year 2006-2007 [64]. (b) the prevalence of ToF per 10,000 live brirths from the PAN registry (red bar) compared other CHD cases. (c) Recrrent risk of ToF in first degree-realtives (d) ToF cases observed slightly more in males compared with females (1.4:1) based on data from PAN registry [64].
D-TGA: dextro-Transposition of the great arteries, cAVSD: complate atrioventricular septal defect, HLHS: hypoplastic left heart syndrome, PS: pulmonary stenosis, AS: aortic stenosis, CoA: coarctation of aorta, TOF: tetralogy of Fallot, ASD:atrial septal defects, VSD: ventricular septal defects.

Genetic counselors use empiric risk figures to calculate recurrence risks (RR) for subsequent pregnancies for couples with a child with ToF. The relative risk of ToF in first-degree relatives varies depending on their relationship to the affected member of the family or whether there are multiple affected individuals in the same family (Figure 3-1). For example, if both parents are healthy and

non-consanguineous, the RR when one child is already affected by CHD is low (2-3%) but almost triples when two or more siblings are affected (8%). On the other hand, when the mother or the father is affected, the RR is around 2-5% and 1-2%, respectively [29, 39, 301].

3.1.3 Embryology and anatomy of Tetralogy of Fallot

Tetralogy of Fallot has been classified as an obstructive lesion of the right side of the heart. To understand how the structural components of ToF arise, I will illustrate the normal anatomy of the right ventricle (RV) followed by the anatomical features of ToF and then describe the main embryological events related to ToF anatomical features.

The main function of the right side of the heart is to pump deoxygenated blood to the lungs. The right ventricle (RV) forms a major portion of the anterior surface of the heart as it extends from the right atrium to the apex of the heart. Traditionally, the RV has been divided into two components: the sinus (inflow) and the conus (infundibulum). The inflow portion extends from the tricuspid valve (TV) to the trabeculated (apical) portion of the ventricle while the outflow portion starts and extends to the pulmonary valve (PV) (Figure 3-2).

ToF is defined by four anatomical features: pulmonary stenosis, ventricle septal defect, overriding of the aorta, and hypertrophy of the right ventricle (Figure 3-3). These four features are thought to arise from a displacement of a single anatomical structure know as muscular outlet septum or the conal septum [302].

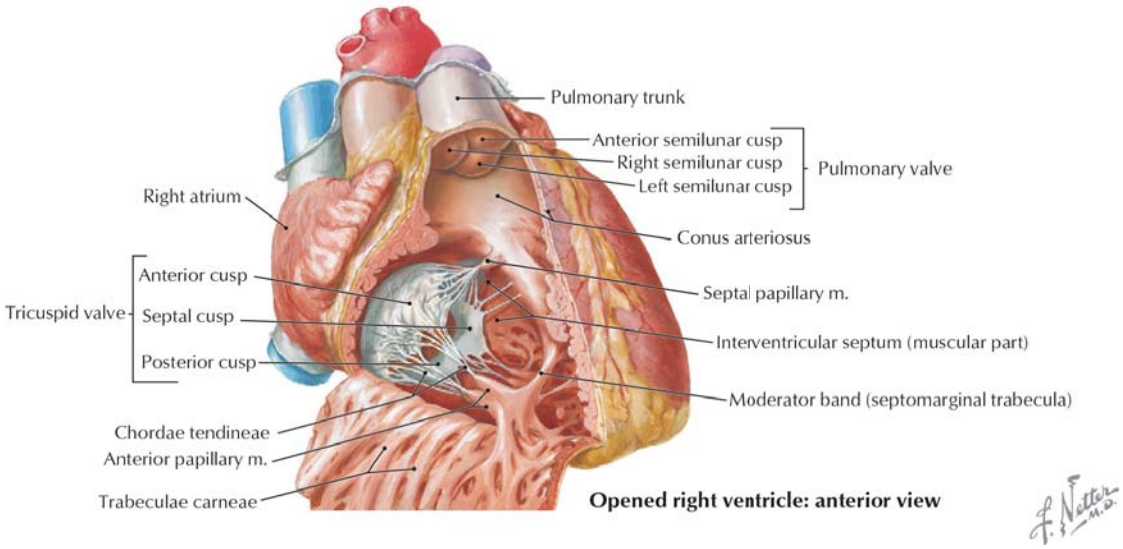


Figure 3-2 The anatomy of the human right ventricle (image adapted from Netter's clinical anatomy [303])

The misalignment of the conal septum narrows the right ventricular outflow tract, leading to subpulmonic obstruction (first ToF feature) and forms a typical misaligment type of ventricular septal defect (second ToF feature). The aortic wall is immediately behind the conal septum so that the left ventricular outflow tract always overrides the misaligned VSD (third ToF feature). Finally, the RV hypertrophy is considered a mechanical consequence of the RV obstruction.

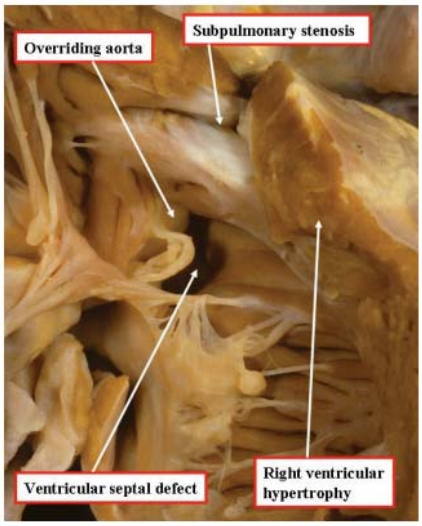


Figure 3-3 The main anatomical features in tetralogy of Fallot (image adapted from [304])

During embryogenesis, these structural abnormalities arise as a result of abnormal development of the outflow tract (OFT) septation. As part of the transition from the heart tube stage to a four-chambered heart, the heart requires proper septation of the outflow tract into the right and left ventricles that open into separate pulmonary and aorta trunks. OFT septation requires multiple cell lineages to participate in cushion growth. For example, neural crest cells (NCCs) migrate into the distal OFT (Figure 3-4-A) and help to develop two groups of cushions: the conal and truncal cushions (Figure 3-4-B,C).

The distal (truncal) cushions fuse to form the aortopulmonary septum, dividing the distal part of the OFT into the aorta and pulmonary trunks [305] while the conal cushions merge to form the conal septum and separating the right and left ventricles [306]. Misaligned or incomplete OFT septation (Figure 3-4-D) leads to a number of congenital heart defects beside ToF such as double-outlet right ventricle (DORV) and transposition of great arteries (TGA) [307].

Up to 16% of ToF cases are associated with other structural or vascular lesions that can influence the clinical presentation of ToF patients and may complicate surgical intervention [302]. The most commonly associated structural lesions are aortic root dilation (40%), peripheral pulmonary stenosis (28%), aortic arch anomalies (25%) and secundum atrial septal defects (20%). Vascular lesions may also accompany ToF, most of which are coronary anomalies (15%), left superior vena cava (11%) or aortopulmonary collaterals (10%) [308].

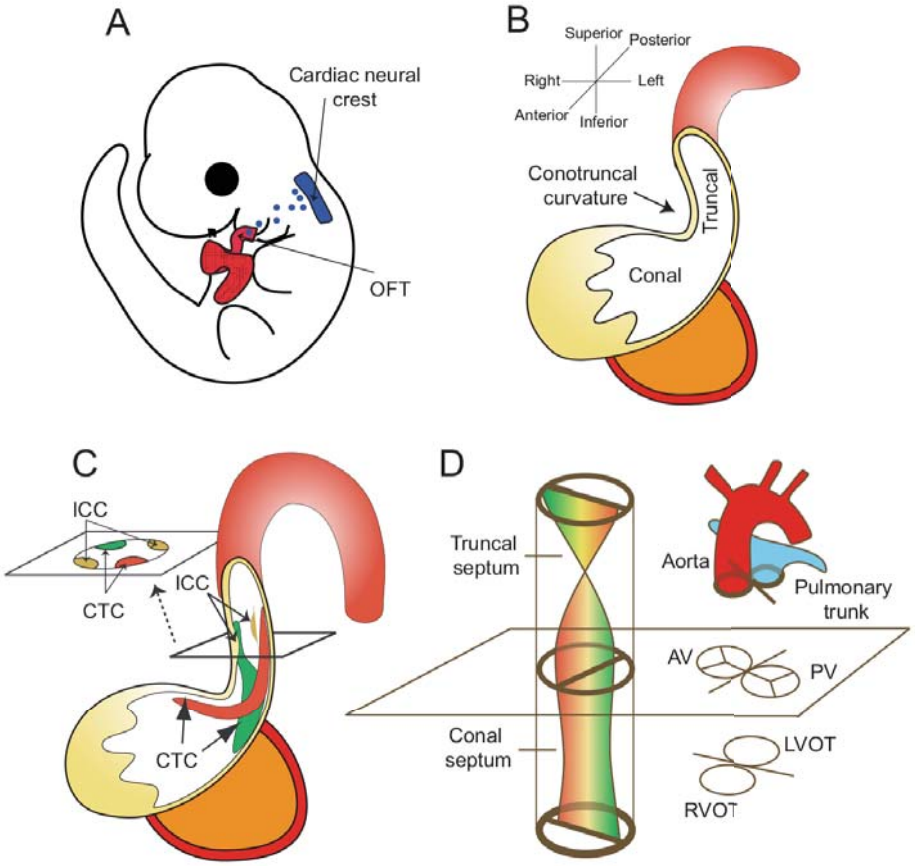


Figure 3-4 Septation of the cardiac outflow tract. (A) Left lateral view of an E10 mouse embryo. The neural crest gives rise to cells (blue) that migrate to and colonize the distal cardiac outflow tract (OFT). (B) The cardiac OFT contains conal (proximal) and truncal (distal) cushions. The boundary between the conal and truncal cushions is marked by an outer curvature of the OFT (the conotruncal curvature). (C) The conotruncal cushions (CTCs) and intercalated cushions (ICCs) develop within the OFT. These cushions occupy four quadrants of the OFT (shown in cross-section). The conotruncal cushions fuse to septate the OFT, as shown in D. (D) Fusion of the conotruncal cushions forms a spiral septum, the truncal part of which divides the OFT into aorta and pulmonary trunk, whereas the conal part septates the OFT into left and right ventricular outlets (LVOT, RVOT). The aortic valve (AV) and pulmonic valves (PV) develop at the conotruncal junction. (Image and caption adapted from [307])

3.1.4 Causes of Tetralogy of Fallot

As for other CHD subtypes, both environmental and genetic causes have been proposed for ToF, and supporting evidence for both is discussed below.

Non-genetic causes

Many environmental factors have been found to increase the risk of the ToF. For example, maternal illnesses during pregnancy such as untreated

phenylketonuria increases the risk of any CHD including ToF > 6-fold, pregestational diabetes (3.1-18 fold), and febrile illness (1.8-2.9) fold [299].

Besides maternal illness, external factors have also been found to increase the relative risk of ToF such as the exposure to organic solvents [9] or carbon monoxide in the first 3-8 weeks of pregnancy [309].

Known genetic causes in syndromic ToF (Mendelian)

Almost 32% of ToF cases occur as part of syndromes with extracardiac phenotypes [310]. The underlying genetic causes of these syndromes range from whole chromosome lesions to single point mutations. Many chromosomal trisomies are associated with ToF. Down syndrome (trisomy 21) has a prevalence of 1 in 700 live births where 44% exhibits various CHD such as complete VSD in 43% and ToF in 6% of the cases [311]. Other trisomies such as Patau syndrome (trisomy 13) and Edwards syndrome (trisomy 18) may present with ToF features [312, 313].

Submicroscopic chromosomal rearrangements may also cause syndromic ToF. The most common submicroscopic chromosomal lesion is 22q11.2 deletion syndrome (1 in 4000 live births), which causes a spectrum of phenotypes ranging from DiGeorge to Shprintzen (velocardiofacial) syndrome wherein CHD are found in 75% of cases [314, 315]. This microscopic deletion spans a 1.5 to 3-Mb region and includes 30-40 genes. One of them is *TBX1*, a known haploinsufficient gene that is likely to be a major contributor to the heart phenotypes [316].

Other genes such as *JAG1* and *NOTCH2* cause Alagille syndrome when they carry point mutations or small insertion/deletion (indel) and exhibit similar clinical symptoms to the 22q11.2 deletion [317]. Alagille syndrome is an autosomal dominant heterogeneous hepato-caridac syndrome where 90-96% of the patients exhibit various CHD [317, 318]. The most common heart defect is pulmonary stenosis (67%) while ToF occurs in 7-16% of the patients [318].

About 89% of the cases are associated with point mutations in the *JAG1* gene, a ligand for NOTCH receptors, while mutations in *NOTCH2* are found in 1-2% of the cases [319]. 50-70% of the mutations in Alagille cases arise *de novo* [319]. The majority of these mutations (~80%) are protein-truncating mutations (frameshift, nonsense, splice site), 7% are whole gene deleting and the remaining are missense mutations [320]. However, some individuals with *JAG1* mutations may express only some of the features of Alagille syndrome, mainly isolated cardiac defects [321-324]. The molecular analysis performed by Fengmin Lu *et al.* [325] in a family with *JAG1* missense mutation that co-segregates with heart defect in absence of liver disease demonstrated a 'leaky' mutation. The leaky mutation affects the amount of Jagged1 protein produced to fall between that seen in an individual with haploinsufficiency and an individual with two normal copies of *JAG1*. The authors suggested that the heart is more sensitive to *JAG1* dosage than the liver.

More recently, specific mutations in the last exon of *NOTCH2* has been shown to cause Hajdu-Cheney syndrome, an autosomal dominant disorder which causes focal bone destruction, osteoporosis, craniofacial dysmorphology, renal cysts, cleft palate, and cardiac defects [326]. These mutations are predicted to disrupt the intracellular PEST (proline-glutamate-serine-threonine-rich) domain and decrease clearance of the notch intracellular domain, thus increasing Notch signalling [326-328]. These findings suggest a complex genotype-phenotype relationship may exist by which different mutations in the same gene can cause completely different monogenic syndromes.

CHARGE syndrome (which stands for coloboma, heart defect, atresia choanae, retarded growth and development, genital hypoplasia and ear anomalies) is another example of a syndrome where 84% of the cases have CHD phenotypes, including ToF in 33% of the patients, and is usually caused by point mutations in the *CHD7* gene [305, 329].

Known genetic causes in non-syndromic ToF

Few genes have been associated with isolated ToF (Table 3-1). Most are based on candidate gene re-sequencing studies. These studies are usually small (< 200 patients) and can explain a small percentage of the cases (~4% on average). Among these candidate genes is *NKX2.5* gene; a transcription factor that is expressed in cardiac mesoderm and its null knockout mouse model halts the heart development at the linear tube stage [330]. Mutations in *NKX2.5* have been found in 1-4% of ToF cases [331, 332] but these two studies did not provide functional evidence to support the effect of these mutations. Other studies confirmed the effect of mutations found in isolated ToF cases by functional studies such as luciferase assays, gene expression and protein localization, modelling mutations in zebrafish (Table 3-1). The strength of evidence from supporting functional experiments varies between studies, which makes establishing genotype-phenotype correlation more difficult.

Table 3-1 Gene mutations in selected candidate genes in isolated ToF from resequencing studies [294]

Gene	Mutated patients / analyzed patients	%	Functional studies	Reference
NKX2.5	6/150	4	N/A	Goldmuntz *et al.* [332]
	9/201	4.5	N/A	McElhinney *et al.*[331]
FOG2	2/47	4	Repression assay	Pizzuti *et al.*[333]
CITED2	3/46	6	Transcriptional assay	Sperling *et al.* [334]
NODAL pathway	15/121	12	Zebrafish rescue assay	Roessler *et al.*[335]
JAG1	3/94	3	Notch activation assay	Bauer *et al.*[321]
	2/112	2.7	N/A	Guida *et al.*[336]
TBX1	3/93	3	Luciferase assay	Griffin *et al.* [337]
FOXA2	4/93	4	N/A	Topf *et al.*[338]
GJA5	2/178	1	Zebrafish modeling and dye transfer studies	Guida *et al.* [339]
FOXC1	1/93	1	N/A	Topf *et al.*[338]
HAND2	1/93	1	N/A	Topf *et al.*[338]

Beside point mutations as a cause of isolated ToF, several recent studies have demonstrated an excess of rare and *de novo* copy number variants (CNV) in non-

syndromic ToF [122, 340, 341]. Greenway *et al.* [341] detected 11 *de novo* CNVs in 114 isolated ToF cases that are novel or extremely rare in 2,265 controls. Some of these CNVs overlap with genes known to cause CHD such as *NOTCH1* and *JAG1*. Based on these findings, the authors predicted that 10% of non-syndromic ToF cases result from *de novo* CNVs. A more recent work by Soemedi *et al* [122] confirmed the burden of large rare genic CNVs in isolated ToF cases but reported a lower rate of *de novo* CNVs in ToF (5%) compared with Greenway *et al*. Silversides et al [340] were able to replicate previous locus-specific findings, such as 1q21.1 deletion CNVs in ~1%, but they also detected CNVs overlapping *PLXNA2* and highlighted the possible involvement of PLXNA2-semaphorin signaling in the development of ToF. The results from the CNV analyses suggest the involvement of novel and multiple genes and pathways in the development of the heart.

At the other end of the spectrum, the "common variant common disease" (CVCD) hypothesis proposes that co-occurrence of multiple common variants, each with a small effect size, is required to cause a complex disease [342, 343]. Genome-wide association studies (GWAS) using SNP arrays have detected hundreds of common variants associated with many complex diseases (a full-catalogue of these studies is available in [114]). Because GWAS requires large sample sizes to detect strongly significant modest effect sizes at the genome-wide level, few studies have detected such signals in CHD. Very recently, Cordell *et al.* [344] published the first example of a GWAS of a CHD subtype (ToF). The authors detected a region on chromosome 12q24 in a northern European discovery set of 835 ToF cases and 5,159 controls (P=1.4 × 10^{-7}) and were also able to replicate the signal in 798 cases and 2,931 controls (P=3.9 × 10^{-5}). The strongest signal detected was for rs11065987, a marker located on 12q24 that had previously been associated with other complex conditions including celiac disease [345], coronary artery disease [346] and rheumatoid arthritis [347]. The strongest candidate gene within the 12q24 region is *PTPN11*, a regulator of Ras/mitogen-associated protein kinase signaling. Mutations in *PTPN11* are a known cause of Noonan's syndrome in which malformation of the cardiac outflow tract is a typical feature [348]. This study also identified a few

interesting signals in other genes such as *GPC5*, a gene encoding glypican 5, which belongs to a family of genes known to work as regulators in many developmental signaling pathways, including the Wnt, Hedgehog, fibroblast growth factor and bone morphogenetic protein pathways [349].

3.1.5 Aim of the study

The aim of this project is to detect genes significantly enriched for rare and / or *de novo* coding variants in isolated ToF cases using a trio-based study design based on exome sequencing.

3.1.6 Overview of the ToF analyses

The molecular genetic studies of ToF, described above, paint a picture of a broad spectrum of aetiologies that range from monogenic forms of ToF at one end to environmental risk factors and common susceptibility variants (multifactorial) at the other. I decided to use exome sequencing to identify highly penetrant coding variants. I used a two-stage study design, with an initial discovery phase using exome sequencing of parent-offspring trios to identify candidate genes, and then a second phase of custom targeted sequencing of these candidate genes in a much larger set of patient-offspring trios (n=250).

In analyzing these data, first I tried to identify genes with plausibly pathogenic *de novo* mutations or inherited variants under autosomal recessive and X-linked models. I also tried to identify genes enriched for inherited variants of incomplete penetrance using a modified version of the transmission disequilibrium test (TDT) that I developed and implemented.

I also investigated whether it might be possible to identify a digenic mode of causation whereby rare coding variants in two functionally related genes would be pathogenic. Digenic inheritance (DI) is the simplest form of inheritance when we consider polygenic disorders. Five decades ago, Defrise–Gussenhoven discussed the subject of reduced penetrance under the monogenic model and suggested that a two-locus model could explain the inheritance more accurately [350]. Currently, there are tens of syndromes that show DI but only a few have

been successfully replicated with supporting evidence from functional studies and / or animal models [351]. Alejandro Schäffer has provided an operational definition of DI: '*inheritance is digenic when the variant genotypes at two loci explain the phenotypes of some patients and their unaffected (or more mildly affected) relatives more clearly than the genotypes at one locus alone*' [351].

The most well studied example of DI is retinitis pigmentosa, which was also the first example of DI in 1994 based on the analysis of multiple pedigrees [352]. Most of the DI studies used either candidate genes design or genetic linkage design [351]. The massively parallel sequencing (MPS) platforms have the potential to facilitate both DI study designs, because they are able to screen all known genes in every sample in the study. To date, only two DI studies used MPS: the first was facioscapulohumeral muscular dystrophy (FSHD) type 2 [353] and the second ataxia and hypogonadism [354]. This analysis is discussed in section 3.3.3 in this chapter.

Finally, I investigated whether I could detect an enrichment of rare coding variants in distinct pathways and this is discussed in section 3.3.4. Additionally, next generation sequencing data can also be used to detect copy number variants, which I discuss in section 3.3.1.4.

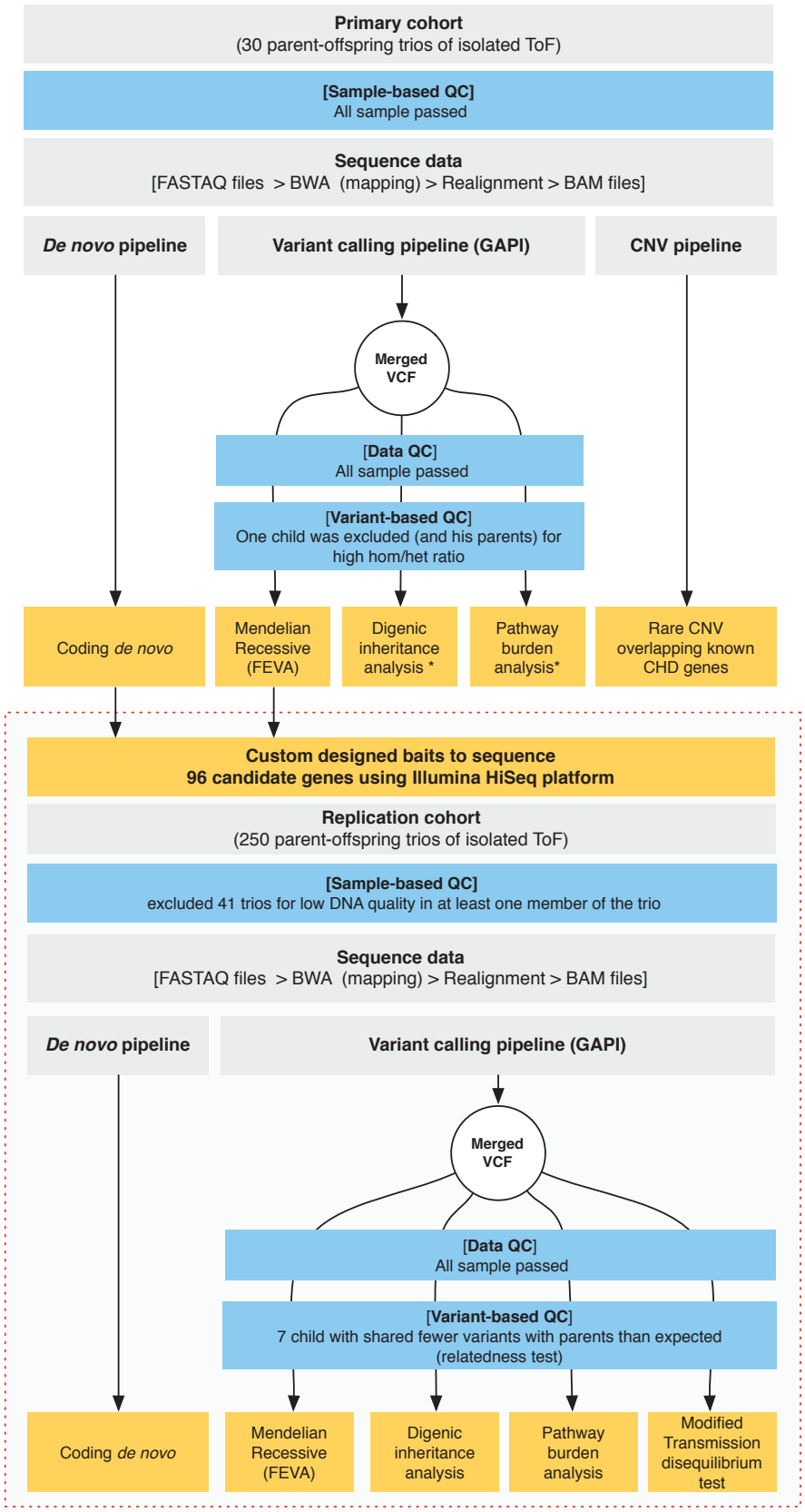


Figure 3-5 A two-stage study design was adapted in this chapter. The first stage included exome sequencing of 30 parent-offspring trios isolated ToF while the second stage included an additional 250 trios as a replication cohort (red dashed box). Quality control (QC) tests (blue boxes) helped to exclude trios that performed poorly on QC test at the level of samples (DNA), data or variant calls. Various analytical approaches (orange boxes) are described in the results section.

*Indicates tests performed after designing the custom baits and thus any identified candidate genes in those tests was not included in the replication cohort. FEVA: Family-based Exome Variant Analysis

3.2 Methods

Samples and inclusion criteria

The primary cohort includes 30 trios of Tetralogy of Fallot children and their healthy parents. These trios are part of the CHANGE cohort managed by Bernard Keavney and Judith Goodship at Newcastle University. The diagnosis was confirmed by echocardiography and only isolated non-syndromic cases were included. The replication cohort of 250 trios of ToF was also selected from the CHANGE cohort using the same inclusion criteria.

Exome sequencing

Samples were sequenced at the Wellcome Trust Sanger Institute. Genomic DNA from venous blood or saliva was obtained and captured using SureSelect Target Enrichment V3 (Agilent) and sequenced (HiSeq Illumina 75 bp pair-end reads). Reads were mapped to the reference genome using BWA [149]. Single-nucleotide variants were called by SAMtools [272] and GATK [153] while indels were called using SAMtools and Dindel [158]. Variants were annotated for allele frequency using 1000 Genomes (June 2012 release) [155] and 2,172 healthy parents from the Deciphering Developmental Disorders project (DDD) [260]. The Ensembl Variant Effect Predictor [170] was used to annotate the impact on the protein structure.

Validation with capillary sequencing

For samples with limited DNA, my colleague, Sarah Lindsay, amplified the whole genome using illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, USA). I used BatchPrimer3 server [355] to design the PCR primers with the default settings. Dr. Lindsay performed the variant validation using capillary sequencing (Genetic Analyzer from Life Technologies, USA). DNA sequences were aligned to the genome reference and analyzed using Geneious Pro (version 5.4.6) [356].

3.3 Results

3.3.1 DNA samples

The primary dataset comprises exome sequences for 30 complete trios of children diagnosed with Tetralogy of Fallot and their healthy parents (all Caucasian). The DNA samples were provided by Professor Bernard Keavney and Judith Goodship from the University of Newcastle. None of the selected patients in this cohort have any other extra cardiac symptoms upon clinical examination. The definitive final diagnosis of the heart defect was confirmed by echocardiography.

3.3.1.1 Quality Control

In order to obtain a high quality dataset for downstream analysis, several quality control assessments are required to detect issues such as contamination, sample swapping or failed sequencing experiments. DNA quality control is applied prior to exome sequencing and data quality control is applied after exome sequencing at the level of both the sequence data (BAM files) and the called variants (VCF files).

DNA quality control

The sample logistics team at the Wellcome Trust Sanger Institute tested the DNA quality of each sample using an electrophoretic gel to exclude samples with degraded DNA. The team also assessed DNA volume and concentration using the PicoGreen assay [277] to make sure every sample met the minimum requirements for exome sequencing. Additionally, 26 autosomal and four sex chromosomal SNPs were genotyped as part of the iPLEX assay from Sequenom (USA). This test helps to determine the gender discrepancies, relatedness or possible contamination issues. All trios in the primary cohort for exome sequencing (30 trios) passed these tests.

Sequence data quality control

The second group of quality control tests was performed once the sequence reads had been generated by the next-generation sequencing platform. Carol Scott at the Genome Analysis Production Informatics (GAPI) team performed these tests to detect samples with too low sequence coverage. None of the trios in the primary cohort failed any of these assessments. The average sequence data generated per exome was 6.2 Gb with 68-fold mean depth and 88% of the exome covered by at least 10 reads.

DNA variant quality control

The third phase of quality control assessed the called variants in the Variant Call Format (VCF) files [161]. The aim of these tests was to detect any outlier samples based on the counts of single nucleotide variants (SNV) or insertion/ deletion variants (INDEL) in comparison to other published and/or internal projects (Table 3-2). All 90 samples in the primary cohort (30 complete trios) showed comparable QC matrices to other internal projects except one sample (TOF5136022) that showed a high heterozygous–to-homozygous ratio ~3.0 instead of the average ratio of ~1.5. This is often a sign of possible contamination and was confirmed later by the sample logistic team. This sample was excluded from the downstream analysis along with its parents. The average numbers of rare and common variants in different classes such as loss-of-function, functional, silent are listed in Table 3-2 and Figure 3-6 and Figure 3-7. All of the QC parameters of the remaining samples are comparable to other internal projects.

Table 3-2 Average counts of various quality matrices and variants classes per sample.

Phase	Goals	Measures	Average per sample
Exome sequencing	Base-level stats	Raw output	6.2 billion
		High quality bases > Q30	88%
		Average coverage per base	68
	Read-level stats	Raw read count	82 million
		Duplication fraction	11%
		High quality mapped reads	62 millions
Variant calling	Single nucleotide variants (SNVs) stats	Total number of coding SNVs	21,367
		Transition/Transversion ratio	3.02
		Het/hom ratio (all coding variants)	1.62
		% Of common coding SNVs (MAF > 1%)	95.4%
		Common loss-of-function variants	80
		Common functional variants	9,629
		Common silent variants	10,271
		% Of rare coding SNVs (MAF< 1%)*	4.5%
		Rare loss-of-function variants	15
		Rare functional variants	608
		Rare silent variants	325
	Insertion and deletion (indels) stats	Total number of coding indels count	436
		% Of common coding INDELs (MAF > 1%)	86%
		Coding in-frame indels	261
		Coding frameshift indels	175
		Coding in-frame / frameshift ratio	1.49
		Rare coding indels	60

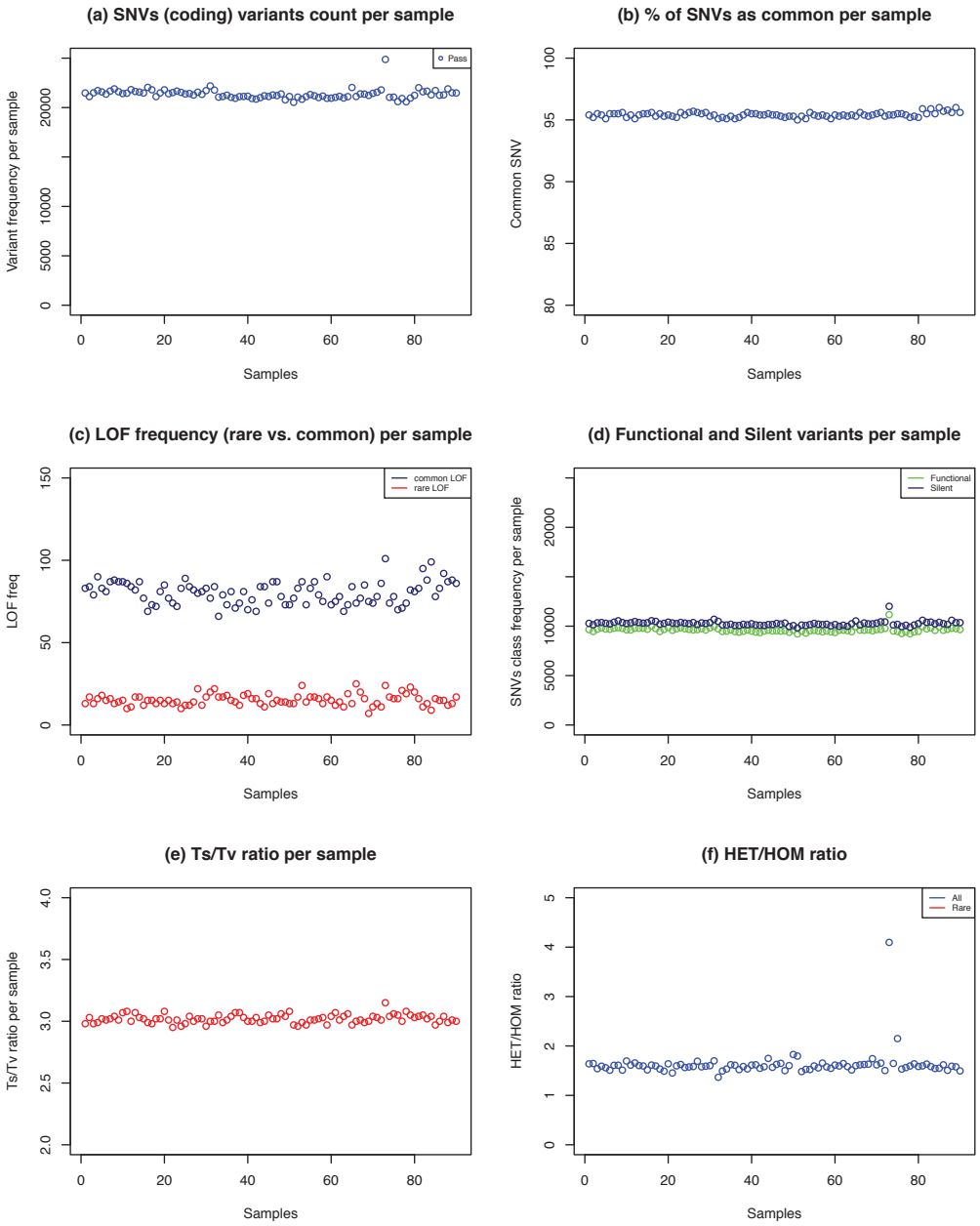


Figure 3-6 Quality control plots including global counts and various single nucleotide variants stats (see main text for description)

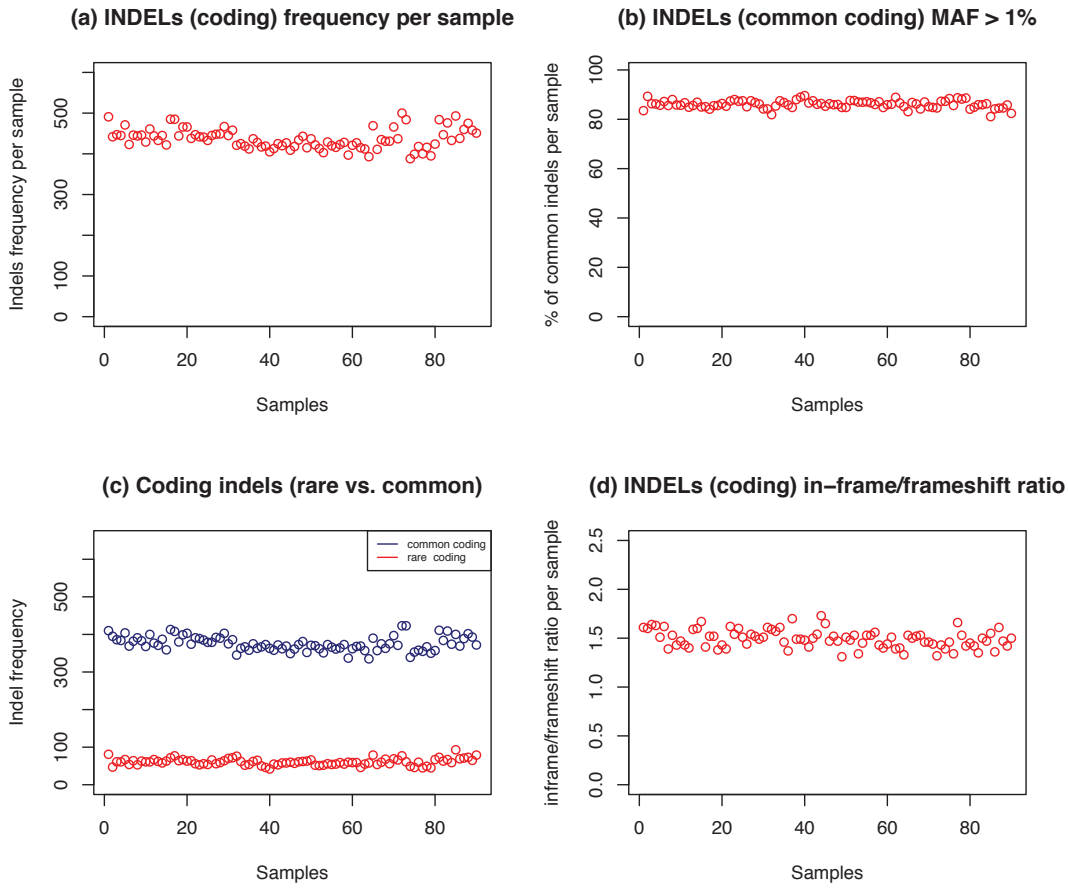


Figure 3-7 Quality control plots for insertion and deletion variants (indels)

3.3.1.2 *De novo analysis (primary cohort)*

The trio study design allows the detection of *de novo* variants. I submitted each trio in the primary ToF cohort to the DeNovoGear pipeline that I developed (described in chapter 2) to detect and annotate candidate *de novo* variants.

Before filtering the DeNovoGear output, each trio had 176 unfiltered candidate *de novo* variants on average (ranges between 113 and 265). However, the raw output was enriched for false positive (FP) variants and thus required stringent filters to minimize the FP rate. I applied five different filters to exclude low quality, non-coding and/or common variants. These filters excluded: (i) variants in tandem repeat or segmental duplication regions, (ii) common variants with minor allele frequency > 1% in the 1000 genomes [155], NHLBI-ESP exome project [199] and the UK10K cohort [264], (iii) when > 10% of the reads in either parent support the variant allele (i.e. the variant is more likely to be

inherited from a parent), (iv) variants not called by an independent caller such as SamTools, Dindel or GATK, and (v) variants predicted to be non-coding and outside canonical splice sites by the VEP annotation tool [170].

Table 3-3 lists the number of filtered candidate *de novo* variants grouped by their predicted effect on the protein structure after applying the above five filters.

Table 3-3 Candidate coding *de novo* variants passed the five filters from 29 ToF trios

Variant predicted consequences	**Count**
Missense	39
Synonymous	8
Splice region	7
Stop gained	6
Frameshift	2
Splice acceptor	2
Splice donor	1
Total	65

To see how these filtered candidate *de novo* variants are distributed in the ToF trios, I plotted the number of variants in each trio in (Figure 3-8). The average number of filtered candidate coding variants per trio is ~2.1. However, three trios did not have any filtered candidate *de novo* coding or splicing variants while only one trio, TOF5135947, showed an excess of filtered candidate *de novo* variants (7 mutations: two loss of functions (stop gain and slice site donor) and five missense variants). The most frequent variant class was the missense (n=39) followed by synonymous (n=8).

Upon validation using capillary sequencing, performed by my colleague Sarah Lindsay, only a third of these variants were found to be true positive while the remaining candidates are either inherited variants, false positive (i.e. reference) or failed sequencing after three attempts (Table 3-4).

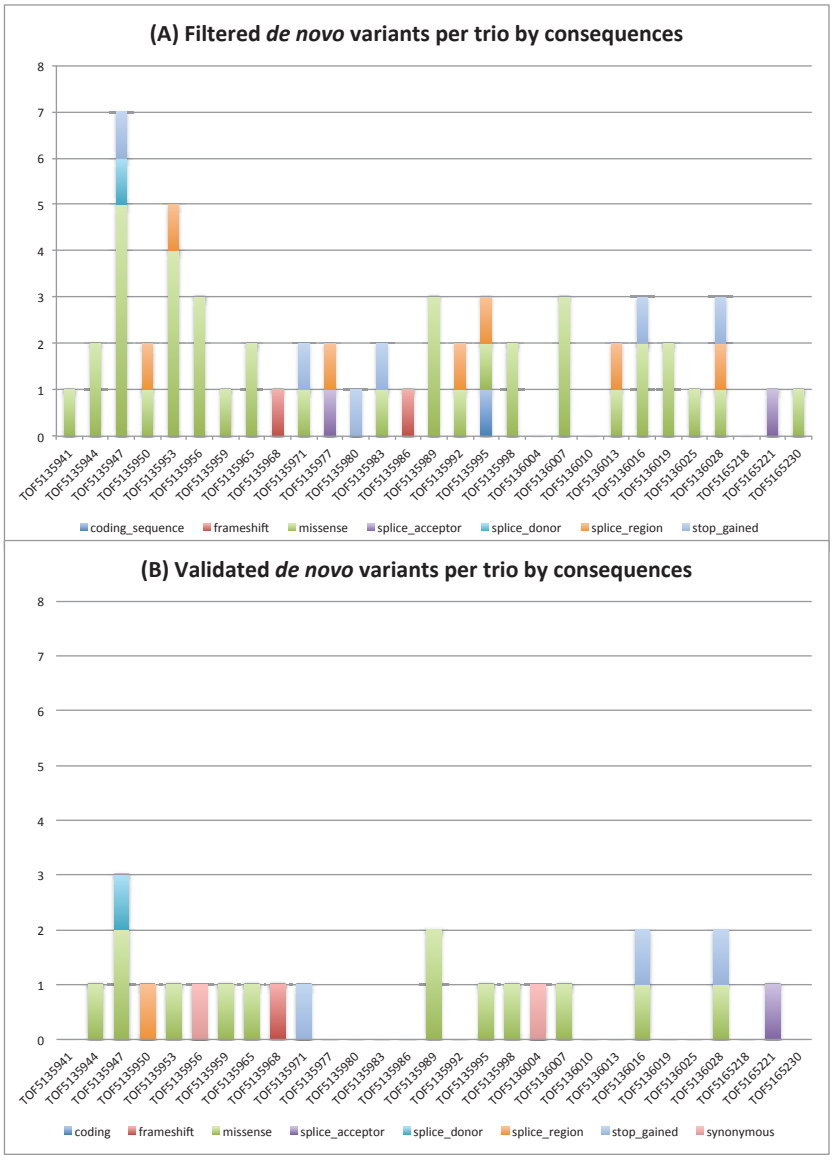


Figure 3-8 Filtered candidate *de novo* variants per trio by consequences. (B) Validated de novo variants by capillary sequencing.

Table 3-4 **Summary of capillary sequencing validation experiment**

Validation results	**Count**
True positive DNMs	21
False positive DNMs	8
Inherited variants	16
Failed sequencing or not enough DNA	19
Pending validation	1
Total	65

The 21 validated coding *de novo* variants are listed in Table 3-5 along with their genome loci and the predicted consequences on the protein structure. The average numbers of SNVs or INDELs in this cohort are comparable to other published studies (Figure 3-9). Excluding INDELs, I observed a significant excess of loss-of-function mutations (~15%) compared to the rate previously reported in controls ~3.4% (exact binomial test P= 0.025) [357] but not for missense (P= 0.06) or splice sites (P = 0.29).

Among the genes with validated *de novo* coding variants, there are three genes known to cause structural heart defects in human and/or knockout mouse models (*NOTCH1, DCHS1* and *SPEN)*.

The *NOTCH1* is the only gene with recurrent *de novo* variants in the primary ToF cohort (a confirmed missense and a single-base deletion predicted to disturb the acceptor splice site of the sixth exon waiting for additional DNA aliquote). *NOTCH1* belongs to a family of four genes encoding single-pass transmembrane receptors that regulate cell fate decisions during development and that are involved in many cellular processes (reviewed in [358]). Dominant mutations in *NOTCH1* have been associated with left ventricular outflow tract abnormalities in human such as coarctation of the aorta, hypoplastic left heart syndrome, bicuspid aortic valve, and aortic valve stenosis [359-361].

The *DCHS1* gene is a member of the cadherin superfamily of cell-cell adhesion molecules and its homozygous knockout mouse model exhibits defects in atrial septation [362]. The third gene with a knockout mouse model showing CHD is *SPEN*. The mouse model died around day 14.5 with morphological abnormalities in the pancreas and heart [363]. However, the *de novo* variant in *SPEN* gene is predicted to be silent and thus unlikely to be causal.

One novel gene in particular worth discussing here is *ZMYM2*, a transcription factor and part of a BHC histone deacetylase complex with a *de novo* coding frameshift [364]. Translocation of this gene with the fibroblast growth factor receptor-1 gene (*FGFR1*) results in a fusion gene, which has been found to cause

stem cell leukemia lymphoma syndrome (SCLL) [365]. This fusion gene was also found to activate the Notch pathway in murine ZMYM2-FGFR1–induced T-cell lymphomas [366]. Although this gene does not have any published knockout mouse model yet, its involvement in the Notch pathway made this gene an interesting candidate for modelling in zebrafish (see zebrafish morpholino knockdown experiments section).

The remaining genes with validated *de novo* coding or splicing variants do not have clear biological links to the development of the heart. Nonetheless, I selected them for re-sequencing in a larger number of samples (see replication study section) to detect any recurrent *de novo* variants in these genes.

Table 3-5 List of validate *de novo* variants from 29 ToF trios. * Pending validation.

Gene	Trio Id	Locus	Reference/Alternative	Consequences
ZMYM2	334	13:20567809	TGG/TG	Frameshift
IKZF1	325	7:50467964	C/T	Missense
TTC18	352	10:75037994	G/A	Missense
MYO7B	367	2:128393882	G/A	Missense
NOTCH1	312	9:139399497	C/T	Missense
DCHS1	382	11:6650724	C/T	Missense
OSBPL10	352	3:31918002	C/A	Missense
FAM178A	333	10:102698379	C/G	Missense
ANKRD11	359	16:89350711	A/C	Missense
ADCY5	318	3:123047511	C/T	Missense
PLCXD1	318	X:209880	G/T	Missense
ATP5G1	330	17:46970784	A/G	Missense
TPRA1	402	3:127298623	C/T	Missense
FLOT2	318	17:27209354	C/T	Disturb donor splice site
PLCG2	319	16:81925070	CTTTT/CTT	Near a splice site (<8bp)
ARHGAP35	335	19:47423379	C/T	Stop gained
SERAC1	402	6:158537270	C/A	Stop gained
ITGB4	382	17:73723777	C/T	Stop gained
SPEN	328	1:16256191	A/G	Synonymous
RREB1	366	6:7230783	C/T	Synonymous
PHRF1	356	11:582022	A/G	Missense
*NOTCH1**	549	9: 139396541	CT/C	Disturb an acceptor splice site

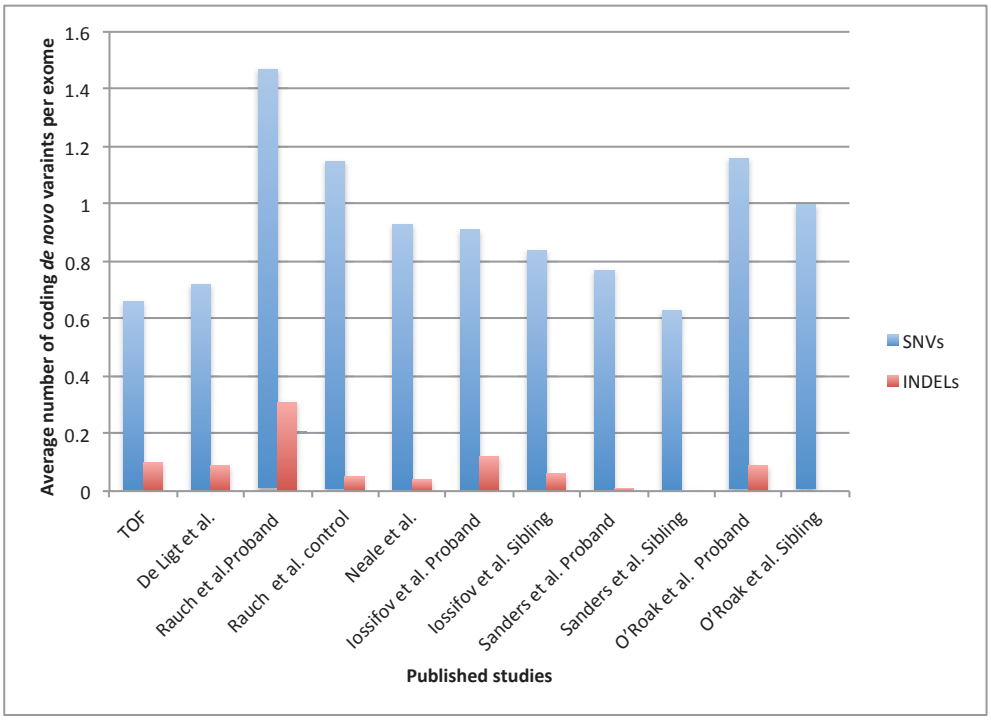


Figure 3-9 The average number of validated *de novo* in the primary ToF cohort is comparable to other published studies [190, 267-271]. (The literature survey is a courtesy of Dr. Matthew Hurles).

3.3.1.3 Analysis of Mendelian inherited variants (primary cohort)

In addition to *de novo* mutations, I set out to identify monogenic candidate genes harbouring rare inherited variants in these trios, under the assumption that both parents do not have CHD and a model of complete penetrance. Only a few inheritance scenarios are compatible with these assumptions. The first scenario is the autosomal recessive model where both parents are heterozygous carriers of the same variant while the child is homozygous. This model can be extended to compound heterozygosity in the child where each allele is inherited from only one parent. The third scenario considers the X-chromosome and is slightly more complex for a few reasons. First, the X chromosome is haploid in males and diploid in females but the variant caller programs (such GATK and SamTools [152, 153]) are not able to differentiate between homozygous or hemizygous status. The second factor to consider is that X inactivation process is random, but can be skewed in some cases [367] which may affect penetrance under an X-linked dominant model. For these reasons, I considered two different scenarios

when dealing with variants on the X chromosome. The first scenario is when a female child inherits an allele from the mother's inactive X-chromosome while the daughter have a skewed X inactivation (Table 3-6**,** B). The second scenario is when a male child inherits an allele from a carrier mother (Table 3-6, C).

I used the Family-based Exome Variant Analysis (FEVA) software that I developed in chapter 2 to output candidate variants for each trio under each scenario. Table 3-6 lists the average number of loss of function (include stop gain, frameshift and variants that disturb either donor or acceptor splice sites), and functional variants (including missense and stop lost). FEVA reported total of ~6.0 rare coding variants per trio regardless of gender. Half of these variants (~2.6 per trio) are autosomally inherited while the rest are inherited on the X-chromosome.

Under these four Mendelian inheritance scenarios, this analysis picked up 159 unique genes with rare coding variants: 51 genes under autosomal recessive homozygous, 58 autosomal recessive compound heterozygous, and 50 genes were X-linked model in either male or female probands.

The vast majority of these candidate genes appear in one sample only except for five genes that appear to be recurrent. All of the five recurrent genes were detected under the compound heterozygous model suggesting that they may be highly variable genes. Based on their biological functions, two out of the five genes (*FLG* and *MUC16*) are less likely to be strong candidates for the ToF or CHD in general. *FLG* encodes a protein aggregates keratin intermediate filaments in the mammalian epidermis while *MUC16* encodes Mucin 16 at mucosal surfaces. The other three genes encode sarcomeric proteins (*TTN*, *NEB* and *OBSCN*) and are known to be very large genes, which may partially explain why they harbor multiple rare coding variants.

Under the X-linked model, four genes appear to be recurrent in female patients only (i.e. variants inherited from the mother). These are *IL13RA1* (interleukin 13 receptor, alpha1), *IRAK1* (interleukin-1 receptor-associated kinase 1), *TLR7* (toll-

like receptor 7), and *ZNF674*. All of these genes, except for *ZNF674*, have knockout mouse models but none show any gross structural heart phenotypes and thus they are unlikely to be strong candidates for ToF [368-370]. *ZNF674* has been linked to nonsyndromic X-linked mental retardation [371] and there is no obvious evidence to support its involvement in heart development.

Table 3-6 Average number of genes with coding variants (excluding silent variants) per offspring in the primary ToF cohort (males=11 and females =18) under different mode of inheritance. The numbers in trio genotype combination column correspond to homozygous reference (0), heterozygous (1), and homozygous non-reference or hemizygous on the X chromosome (2) and are ordered as the child, mother and father, respectively.

Chromosome	Genotypes		Variant type		
	Genotype status	Trio combination	Loss of function	Functional	Both
[A] Autosomal	Homozygous	(2,1,1)	0.03	0.34	0.37
	Compound heterozygous	Locus A (1,1,0) Locus B (1,0,1)	0.07	2.17	2.24
[B] X in females	Heterozygous	(1,1,0)	0.22	3.22	3.44
[C] X in males	Hemizygous	(2,1,0)	0.09	3.55	3.64

3.3.1.4 Copy Number Variant analysis (primary cohort)

Rare copy number variants are known to cause 5-10% of isolated ToF cases [122, 340, 341] based on array CGH and SNP array. Recently, several groups have published computation approaches to call CNVs from exome data (reviewed in [157]). Calling CNV from exome data is still in its infancy and consequently is associated with a relatively high false positive rate. However, I decided to investigate the possibility of *de novo* or rare inherited CNVs that overlap with known CHD genes.

My colleague, Dr. Parthiban Vijayarangakannan, has developed a CNV-calling algorithm and software called CoNVex [372] to detect copy number variation from exome and targeted-resequencing data using comparative read-depth. He generated the CNV calls from the primary ToF cohort and I performed the downstream analysis.

Initially, I was able to detect two plausible *de novo* duplication events in two trios out of 29. The first is a 218Kb duplication on chromosome 2 and spans several genes including *HDAC4* (Histone deacetylase 4). The second CNV event was a 1.6Mb duplication overlapping with the *PFKP, PITRM1*, and *ADARB2* genes (Figure 3-10 and Table 3-7).

HDAC4 encodes a protein with deacetylation activity against core histones [373] and *HDAC4*-null mice display premature ossification of developing bones but did not exhibit heart phenotypes [374]. However, the haploinsufficiency of *HDAC4* causes brachydactyly mental retardation syndrome, which has been associated with cardiac defects in 20% of the patients [375, 376]. Moreover, overexpression of *HDAC4* inhibits cardiomyoblast formation and down-regulate the expression of *GATA4* and *Nkx2-5* [377]. Further investigations are required to determine the dosage sensitivity of *HDAC4* and the nature of its role in heart development.

On the other hand, none of the genes that overlap with the second *de novo* duplication have a knockout mouse model (*PFKP, PITRM1,* and *ADARB2*). The *PFKP* gene encodes the platelet isoform of phosphofructokinase and a key metabolic regulator of glucose metabolism [378]. *PITRM1* is a zinc metalloendopeptidase that has been implicated in Alzheimer's disease and mitochondrial peptide degradation. More recently, the hedgehog signalling was found to regulate *Pitrm1* in the developing mouse limb [379]. The last gene, *ADARB2*, encodes a protein that is a member of the double-stranded RNA (dsRNA) adenosine deaminase family of RNA-editing enzymes [380]. None of these genes have strong evidence to support a direct involvement in heart development.

My aim in the second part of CNV analysis was to find recurrent rare inherited CNV that overlap with known CHD genes in human and/or animal models. To obtain this callset, I applied four filters on the original CNV calls from CoNVex pipeline : (i) CNV calls with CoNVex scores < 10 were excluded to remove low quality calls, (ii) CNV calls with > 50% of their length overlapping known common CNV manually curated from multiple high-quality publications and used

as part of CoNVex pipeline, (iii) I excluded CNV calls with frequency > 1% in CHD samples sequenced by our group (n=723) , and (iv) I excluded CNV calls that do not overlap with candidate CHD genes (n=1,507 genes manually curated from CHD studies in human and animal models, curtsey of Dr. Marc-Phillip Hitz).

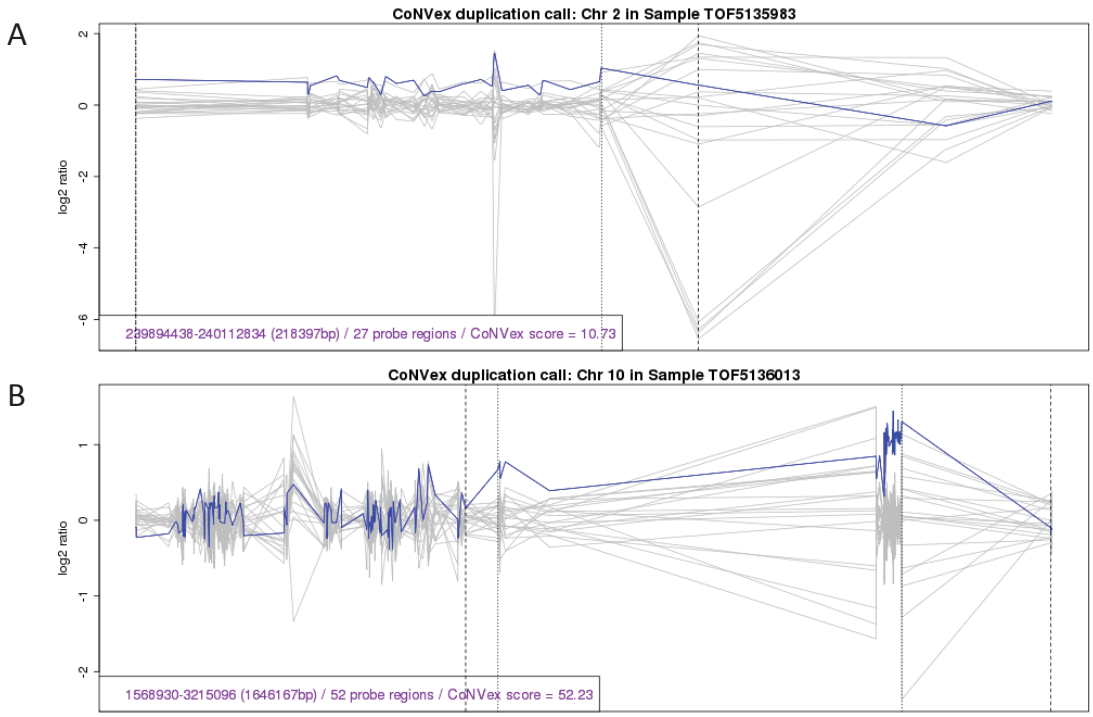


Figure 3-10: (A) A 218Kb duplication event on chromosome 2 spanning the *HDAC4* gene in patient (TOF5135983). The blue line is the log2 ratio in the patient while the grey lines represent the log2ratio scores for the same region in other samples in the cohort. (B) A 1.6 duplication event on chromosome 10 spanning the *PFKP*, *PITRM1* and *ADARB2* genes in patient TOF5136013.

Table 3-7 Plausible *de novo* duplications in the primary ToF cohort. DUP: duplication. Chr: chromosome, Number of probes: number of baits covering CNV. The CoNVex score is a confidence score based on the Smith-Waterman score divided by the square root of the number of probes where higher values mean better and more confidant calls.

Sample ID	Chr	Start	End	Number of probes	CoNVex Score	CNV type	Genes
TOF5135983	2	239894438	240112834	27	10.73	DUP	*HDAC4, MIR4440, MIR4441*
TOF5136013	10	1568930	3215096	52	52.23	DUP	*PFKP, PITRM1, ADARB2*

Only three trios were found to have two small inherited duplications (1.3 Kb, and 12.7Kb) that span *FOXC1* and *FOXC2*, respectively (Table 3-8). *FOXC1* and *FOXC2* are both forkhead box transcription factors crucial for development of the eye, cardiovascular network, and other physiological systems. The mice null models show various structural heart defects [381, 382]. Mutations in *FOXC1* in particular have been associated with aortic stenosis, pulmonary valve stenosis and atrial septal defect [383]. However, it is unlikely to identify the same rare duplication in three unrelated trios in a small sample size and thus these duplications are likely to be false positive. Moreover, the number of probes overlapping these two duplications is small (one or two probes). Validation using an alternative CNV detection method (e.g. custom designed array or MLPA [384]) is required before considering these interesting findings any further.

Table 3-8 List of recurrent rare inherited duplications overlapping known CHD genes. DUP: duplication

Sample ID	Chr	Start	End	Number of probes	CoNVex Score	CNV type	Genes	Inherited from
TOF5135968	6	1610536	1611901	1	13.75	DUP	*FOXC1,*	Paternal
	16	86600787	86613488	2	11.14	DUP	*FOXC2, FOXL1, RP11-463O9.5,*	Maternal
TOF5135971	6	1610536	1611901	1	14.64	DUP	*FOXC1,*	Both parents have this CNV
	16	86600787	86613488	2	13.83	DUP	*FOXC2, FOXL1, RP11-463O9.5,*	Father
	X	153283293	153285567	1	11.34	DUP	*IRAK1, MIR718,*	Mother
TOF5135977	6	1610536	1611901	1	13.22	DUP	*FOXC1,*	Maternal
	16	86600787	86613488	2	10.56	DUP	*FOXC2, FOXL1, RP11-463O9.5,*	Maternal

3.3.2 Replication study

In the second stage of the study I designed custom baits to capture coding regions of 122 candidate CHD genes for sequencing in whole genome amplified DNA from 250 parent-offspring trios with isolated ToF. The main goal of this replication study is to identify additional ToF families with mutations in the same genes identified in the primary cohort analyses. Additionally, I wanted to test the burden of rare coding variants in other known candidate CHD genes from published studies that include linkage analysis, candidate genes, genome wide associations and copy number variant studies.

3.3.2.1 Gene selection for replication study

I selected 122 genes for the replication study using three different classes (Table 3-9). The first class includes genes with validated *de novo* coding variants (e.g. *NOTCH1*, *ZMYM2*, and *DCHS1*) in the 29 trios described above or other candidate genes harbouring rare loss-of-function variants in other ToF samples (e.g. the *GMFG* gene that I found to harbor a homozygous stop gain in three affected siblings with ToF in a different study (see section 2.3.6 FEVA applications in chapter 2). The second group of candidate genes includes genes that have been linked to ToF in humans through genetic evidence from candidate gene sequencing, association, CNV and / or linkage studies. The third group includes genes that are involved in the WNT or NOTCH pathways and have been shown to have a clear structural heart phenotype in mouse knockout models.

The WNT/NOTCH pathways have previously been shown to be enriched for rare and *de novo* CNVs in CHD in general and in TOF cases in particular [122, 341] which make them good candidates for sequencing in replication studies. Because the total number of genes involved in the WNT and NOTCH pathways exceeds the available space within the custom bait design, I had to exclude many genes in a systematic fashion. First, I downloaded the mouse knockout phenotype data from the MGI database [288] and then assigned each gene to one of five different levels based on the type and severity of the CHD phenotype and associated GO terms in the mouse model (see the full workflow of mouse CHD genes selection in Figure 3-11). The complete list of selected genes is available in the Table 3-10.

The bait length is 120 and I used the same baits used to cover the genes from Agilent Technologies; Human All Exon 50 Mb (SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing V4). The baits in this kit have been optimized for all candidate genes I have selected for the replication study, except for the *CFC1* gene. *CFC1* was not covered in the original SureSelectXT kit and I added 2x tiling baits to cover it. I also visually inspected

the bait coverage of the genes using the UCSC genome browser to ensure all coding regions were covered properly.

Table 3-9 The rationale and number of selected candidate genes in ToF replication.

Group of genes	Rationale for selection	Number of genes
From primary cohort (exome)	Candidate TOF genes	12
Known ToF genes	Published ToF candidate genes	20
	Gene-based and genome-wide association studies	11
	Candidate genes from linkage analysis studies	4
	Candidate genes from CNV studies	5
NOTCH/WNT pathways	Notch pathway (with heart phenotypes in MGI)	41
	Wnt pathway (with heart phenotypes in MGI)	36
Total		129 (122 unique)

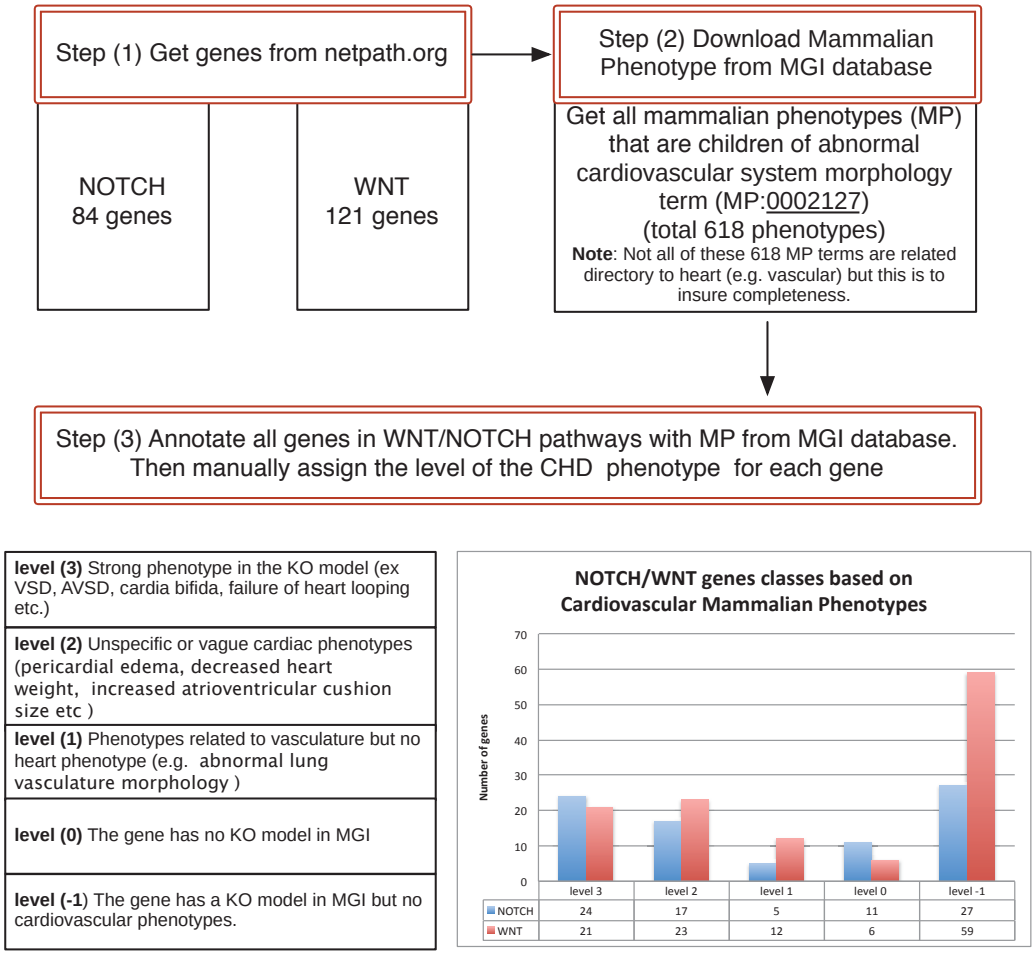


Figure 3-11 The workflow of gene selection from NOTCH/WNT pathway in the ToF replication study

Table 3-10 List of candidate gene selected for the replication study. Some of the candidate genes from primary cohort *de novo* analysis such as *ITGB4* were not included since they were identified after I designed the custom baits.
* The candidate *de novo* variants in *XXYLT2* and *MTUS2* turned out to be false positive during capillary sequencing.
***CFC1* has been covered using tiling probes (1x), while other genes have Agilent's V4 baits that overlap with GENCODE v12.

ADAM10	*ESR1*	*MAP3K1*	*PRKCQ*
ADAM17	*FAT1*	*MAP3K7*	*PSEN1*
ALDH1A2	*FBXW7*	*MAPK1*	*PSEN2*
APC	*FN1*	*MAPK3*	*PTPN11*
APH1A	*FOXH1*	*MAPK8*	*RAC1*
ARHGAP35	*FURIN*	*MEF2C*	*RAF1*
ATR	*FZD1*	*MTHFR*	*RAI1*
AXIN1	*FZD10*	*MTUS2**	*RBPJ*
AXIN2	*FZD2*	*NCOR2*	*RELA*
C2CD3	*GATA3*	*NCSTN*	*ROR1*
CCND1	*GATA4*	*NFATC1*	*ROR2*
CDH18	*GATA6*	*NKX2-5*	*RPS6KB2*
CDH2	*GDF1*	*NODAL*	*SALL4*
CDK2	*GMFG*	*NOTCH1*	*SLC19A1*
*CFC1***	*GPC3*	*NOTCH2*	*SMAD1*
CNOT6	*GPC5*	*NRP1*	*SMAD3*
COL3A1	*HAND2*	*NUMB*	*SPEN*
CRKL	*HDAC1*	*PAX9*	*STAT3*
CSNK2A1	*HDAC2*	*PCDH15*	*TBX1*
CTBP1	*HEY2*	*PCDHB7*	*TBX5*
CTBP2	*IL6ST*	*PCDHB8*	*TCF3*
CTNNB1	*ISL1*	*PCSK5*	*TDGF1*
DAAM1	*JAG1*	*PIK3R1*	*TP53*
DCHS1	*JUN*	*PIK3R2*	*VEGFA*
DLL1	*JUP*	*PLEC*	*VEGFC*
DLL4	*KL*	*POFUT1*	*WNT7B*
DVL1	*LAMP2*	*PPARG*	*XXYLT1**
DVL2	*LPP*	*PPM1K*	*ZFPM2*
DVL3	*LRP5L*	*PRKACA*	*ZMYM2*
EDIL3	*MAML1*	*PRKCA*	
EP300	*MAML3*	*PRKCB*	

3.3.2.2 Quality control (replication study)

Similar to the primary exome sequencing of ToF trios to obtain high quality DNA variants for downstream analyses, different quality control steps were performed at the level of DNA samples, the sequencing data (BAM files) and the called variants (VCF files).

The sample logistics team at the Wellcome Trust Sanger Institute tested the DNA quality of each sample using electrophoretic gel to exclude samples with degraded DNA. The team also tested DNA volume and concentration using the PicoGreen assay [277] to make sure every sample meets the minimum requirements for sequencing. Additionally, 26 autosomal and four sex chromosome SNPs were genotyped as part of the iPLEX assay from Sequenom (USA). These tests excluded 41 out of 250 complete trios submitted for sequencing. The custom sequencing generated 0.35 Gb per sample with an average 267-fold depth within the target regions.

Since the targeted region is much smaller than the regular exome sequence study (122 genes vs. ~20,000 genes in an exome), the basic QC matrices such as the number of variants are expected to be different (Table 3-11, Figure 3-12 and Figure 3-13). However, the transition/ transversion ratio in the replication cohort (~3.3) is comparable to the primary exome-based cohort (~3.1). Similarly, heterozygous / homozygous ratio is also comparable (1.4 in the exome and 1.5 in replication design). On the other hand, the coding in-frame / frameshift ratio is very different (1.5 in the exome and 5.1 in the replication design). This is mainly due to the very low number of indels in the replication design, which is expected given its smaller number of genes. These analyses did not identify any further outlier samples that needed exclusion.

Table 3-11 Quality tests of the exome sequence data and called variants in replication ToF cohort

Phase	Goals	Tasks	Average per sample
Exome sequencing	Base-level stats	Raw output	346 million
		High quality bases > Q30	87%
		Average coverage per base	267
	Read-level stats	Raw read count	4.6 million
		Duplication fraction	25%
		High quality mapped reads	3.2 million
	Single nucleotide variants (SNVs) stats	Total number of coding SNVs	230
		Transition/Transversion ratio	3.34
		Het/hom ratio (all coding variants)	1.72
		% Of common coding SNVs (MAF > 1%)	96%
		Common loss-of-function variants	0.4
		Common functional variants	99
		Common silent variants	121
		% Of rare coding SNVs (MAF< 1%)*	4%
		Rare loss-of-function variants	0.06
		Rare functional variants	4.35
		Rare silent variants	4.41
	Insertion and deletion (indels) stats	Total number of coding indels count	12.4
		% Of common coding INDELs (MAF > 1%)	86%
		Coding in-frame indels	10.5
		Coding frameshift indels	1.82
		Coding in-frame / frameshift ratio	5.11
		Rare coding indels	1.78

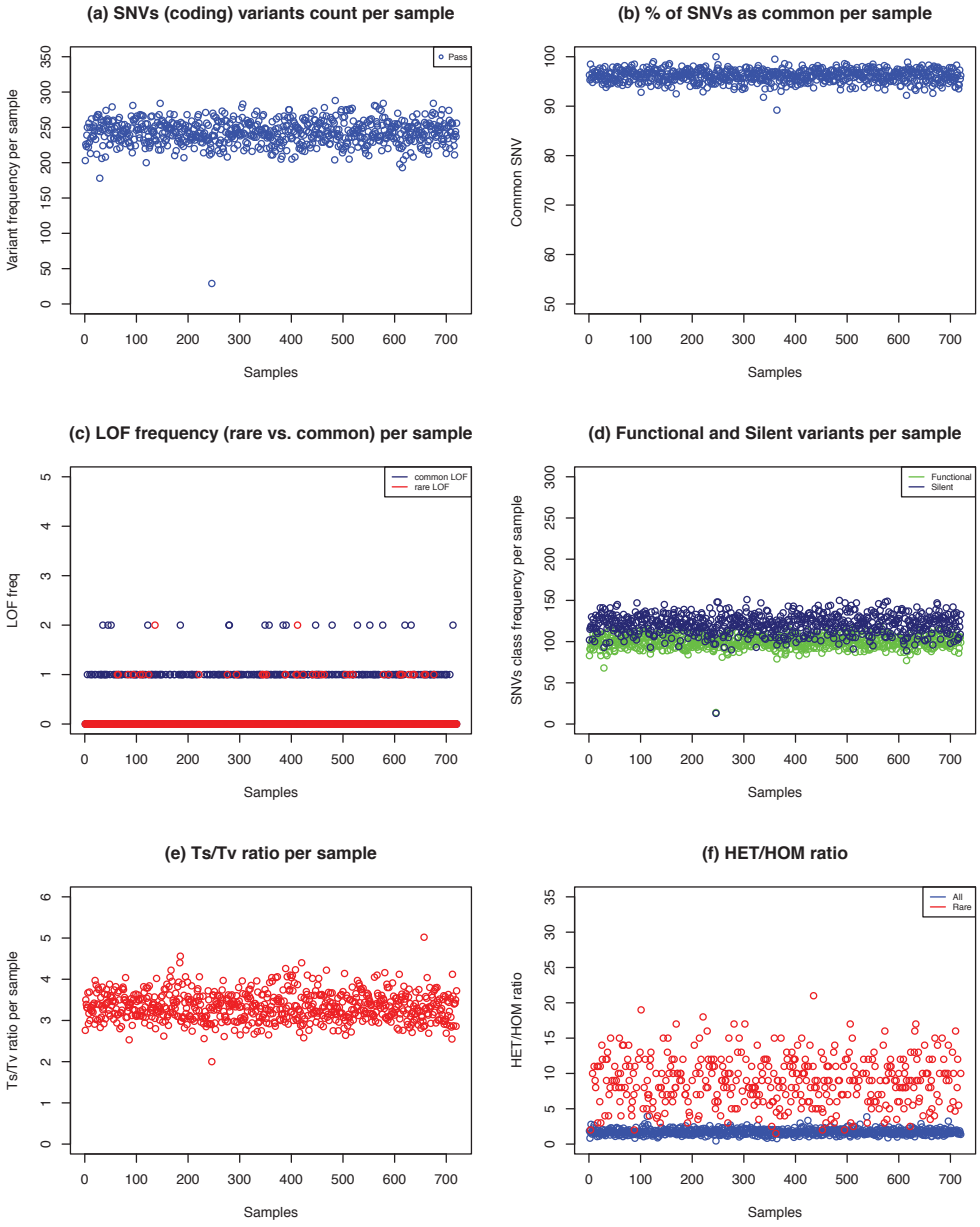


Figure 3-12 Quality control plots including global counts and various single nucleotide variants statistics in 209 trios from the ToF replication cohort (see main text for description)

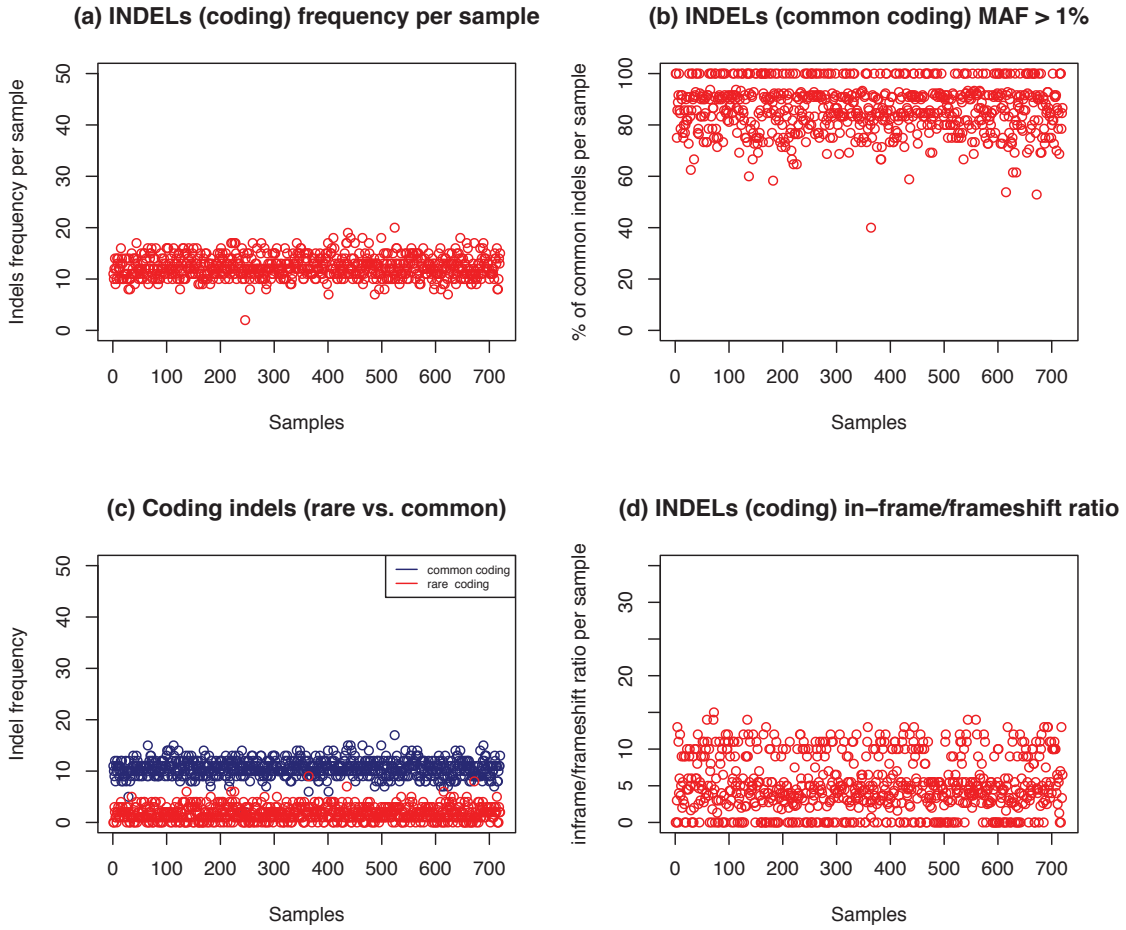


Figure 3-13 Quality control plots for insertion and deletion variants in 209 trios from the ToF replication cohort.

3.3.2.3 *Trio relatedness (replication cohort)*

After performing the sample-by sample quality control tests, I checked trio relatedness *in silico*. My approach was based on examining the number of shared variants between each child and his parents. Most children shared ~71% of their variants on average with each parent (Figure 3-14, red points). To use a control set, I assigned each child to random parents and calculated the percentage of shared variants again (Figure 3-14, blue points) which show children assigned to random parents shared 59% of their variants on average (they mostly share common variants).

I found six outlier samples out of the 209 original trios where each child shared < 62.5% with the father and 65.5% with the mother. The low percentage of shared

variants indicates either a contamination or sample swapping issue. These six samples have been flagged in the downstream analyses in order to spot possible unusual output, but were not excluded from the analysis.

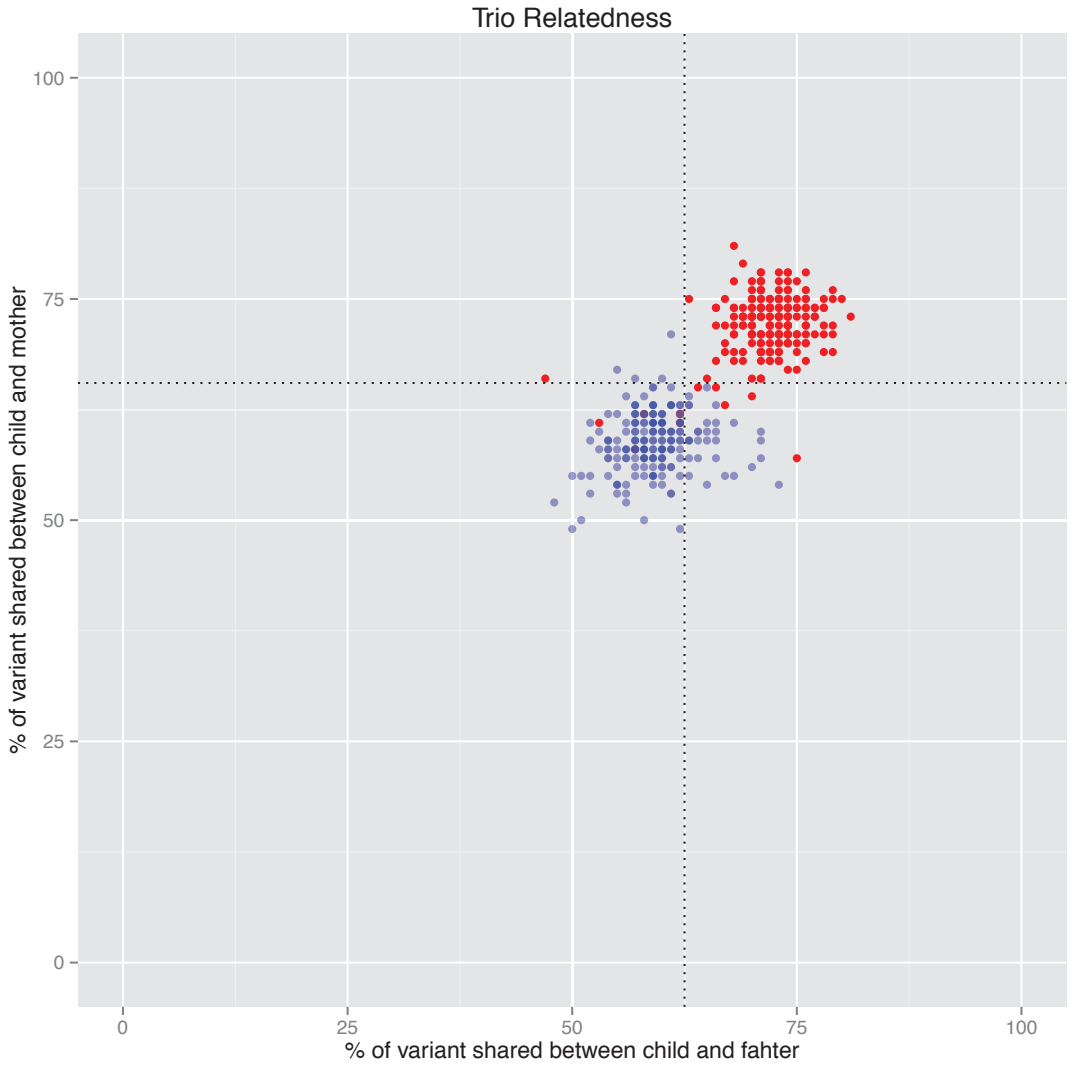


Figure 3-14 Percentage of shared variants between each child and his parents (red) and when children are assigned to random parent pairs (blue). Dashed black lines are used to separate the two groups and to flag six trios where children have shared < 62.5% of their variants with the father and/or < 65.5% with the mother.

3.3.2.4 De novo variant analysis (replication cohort)

The goal of this analysis is to detect *de novo* coding variants in the genes that already have at least one *de novo* coding variant in the primary cohort (Table 3-5).

I submitted all trios to the DenovoGear pipeline I designed (described in chapter 2) and used the same five filters described in the primary ToF cohort to pick coding or splicing rare plausible *de novo* variants that were not seen in the parents and were called by independent programs (GATK, SamTools and/or Dindel). I was able to detect six plausible *de novo* variants in four genes, three of which are loss-of-function (Table 3-12). Two genes had *de novo* mutations in two unrelated trios.

To assess whether the observed number of coding *de novo* variants is more than expected, I calculated the expected number of missense and putative loss of function variants given the cumulative length of coding regions in 122 genes selected for the replication study (329,562 bp), the single nucleotide mutation rate (1.5×10^{-8}), proportion of loss of function (0.052) and proportion of missense (0.663) [357]. In 122 genes from 209 trios, this analysis estimates the expected number of *de novo* missense and loss of function to be 1.3 and 0.1, respectively.

NOTCH1, which already had two *de novo* coding variants in the primary cohort (one missense and one insertion disturbing the acceptor splice site of the 29th exon) had another two plausible *de novo* coding variants in the replication cohort, both of which were loss-of-function (nonsense).

Interestingly, I also detected two plausible *de novo* coding variants in the *JAG1* gene (a missense and a variant predicted to disrupt a donor splice site) that encodes for jagged 1 protein, a known ligand for NOTCH1. Mutations that alter jagged 1 protein have been linked to Alagille syndrome, where 90% of the patients have CHD, mostly right-sided defects ranging from mild peripheral pulmonic stenosis to severe forms of tetralogy of Fallot [317, 318]. The knockout mouse model also showed similarities with Alagille syndrome including various heart defects [385]. However, mutations in *JAG1* have been suggested as a cause for non-syndromic CHD [386] and have also been reported in familial tetralogy of Fallot [323].

The fifth plausible missense *de novo* mutation was detected in *VEGFA*, which encodes for a growth factor that is active in angiogenesis, vasculogenesis and endothelial cell growth. The *VEGFA* mouse knockout model has a delayed and abnormal heart development, including the overriding of the aorta [387, 388]. Moreover, common SNPs in *VEGFA* have been reported to increase the risk of isolated ToF [389].

The last plausible *de novo* missense variants was found in *AXIN1,* which encodes a protein that has both positive and negative regulatory roles in Wnt-beta-catenin signaling during embryonic development and in tissue homeostasis in adults [390]. The homozygotic mouse null model died at embryonic day 8-10, exhibiting neuroectodermal defects and axial duplications. Heterozygotes exhibit underdeveloped trunk, kinky neural tube, enlarged pericardium, and cardia bifida [391]. Moreover, the *axin1* zebrafish (*mbl*) mutants showed an absence of heart looping in 13% of the embryos [392].

Table 3-12 List of plausible *de novo* coding variants that pass quality filters in 209 ToF trios. All variants are missense and predicted by PolyPhen [171] to have a probably damaging (PRD), a possibly damaging (PSD), or a benign (BEN) effect on the protein function. VEP: Variant Effect Predictor [170]. GERP is Genomic Evolutionary Rate Profiling scores where higher values indicate conserved nucleotides) [164].
chr: chromosome, na: not applicable.

Sample ID	Chr	Position	Reference allele	Alternative allele	Gene	Type	Amino acid	PolyPhen	Capillary Sequencing Validation
843	9	139399230	C	T	*NOTCH1*	Stop gained	W/*	Unknown	Confirmed
169	9	139412303	G	A	*NOTCH1*	Stop gained	R/*	Unknown	Confirmed
577	20	10630973	C	A	*JAG1*	Missense	G/W	PRD (0.999)	Confirmed
317	20	10625003	A	C	*JAG1*	Splice donor	na	Unknown	Confirmed
861	16	339545	G	A	*AXIN1*	Missense	A/V	PSD (0.679)	Not validated
780	6	43749703	C	T	*VEGFA*	Missense	P/S	PRD (1)	Not validated

I determined the probability of seeing multiple mutations in the same gene given the size of the gene and the number of patients evaluated in both primary and replication cohorts (Table 3-13). The number of *de novo* variants observed in *NOTCH1* reached genome-wide significant levels for putative loss of function variants (P=3.8 ×10^{-9}) and for missense variants (P=9.4 ×10^{-8}). The number of observed *de novo* mutations in *JAG1* is not significantly greater than the null expectation after applying a Bonferroni correction for multiple testing of 20,000 genes, but it would remain significant after applying Bonferroni correction for multiple testing in the 122 genes in the replication experiment.

Table 3-13 Probability of observing the reported number of *de novo* variant by chance in genes recurrently mutated in this study. The weighted mutation rate is calculated based on the coding gene length, single nucleotide mutation rate (1.5 ×10^{-8}), proportion of loss of function (0.052) or proportion of missense (0.663) [357] and the number of autosomal chromosomes (number of samples × 2=476). The p value is bases on the Poisson distribution density function.

Gene	Captured length (bp)	Variant type	Weighted mutation rate	*De novo* mutation	*P* value †
NOTCH1	7,668	LoF	0.0028	3	3.8 ×10^{-9} ***
		Functional	0.0362	4‡	9.4 ×10^{-8} **
JAG1	3,657	LoF	0.0013	1	0.00135
		Functional	0.0173	2‡	0.00017

† Adjusted α is equivalent to 0.05/20,000 = 2.5 ×10^{-6} (*), 0.01/20,000 = 5.0 ×10^{-7} (**) and 0.001/20,000 = 5.0 ×10^{-8} (***)
‡ Functional de novo variant count include both loss of function and functional *de novo* variants.

3.3.2.5 Mendelian-based variant analysis (replication cohort)

Similar to the Mendelian-based variant analysis in the primary cohort, I generated a list of rare inherited coding and splicing variants under autosomal recessive and X-linked models assuming healthy parents (for more details see Mendelian-based variant analysis in the primary cohort section above).

I defined rare variants as having a minor allele frequency of less than 1% in both the 1000 genomes project data [155] and also in ~2,172 healthy parents from the Deciphering Developmental Disorders (DDD) project [260]. In these analyses I only included variants annotated by the VEP software [170] as being stop gain, frameshift, missense, stop lost or disrupting donor or acceptor splice sites.

I used the family-based variant analysis program (FEVA) to detect 11 candidate genes with rare coding variants under different inheritance models (Table 3-14). Three genes out of 11 appear in more than one trio. The first recurrent gene is *PCSK5* (proprotein convertase subtilisin/kexin type 5) wherein the same frameshift variant appears homozygously in three different samples under an autosomal recessive model (Table 3-15). *PCSK5* belongs to a proconvertase family, which cleave latent precursor proteins into their biologically active products and has been found to mediate post-translational endoproteolytic processing for several integrin alpha subunits [393]. The knockout mouse model exhibited multiple cardiac defects, including atrial and ventricular septal defects [394].

PLEC is the second gene with recurrent rare coding variants under the autosomal recessive compound heterozygous model. One of the patients carries four rare missense variants (one inherited from the father and the other thee from the mother (Table 3-16). *PLEC* encodes plectin-1, an intermediate filament-binding protein, to provide mechanical strength to cells and tissues by acting as a crosslinking element of the cytoskeleton [395]. Plectin-1 is considered to be one of the largest polypeptides known (500-kD). Mutations in this gene have been linked to epidermolysis bullosa simplex [396], while recessive mutations were found in three patients with limb-girdle muscular dystrophy without skin abnormalities [397]. The mouse knockout model did not show gross structural defects in the heart although the histological sections of the heart tissues showed cardiomyocyte degeneration and misaligned Z-disks [398].

The last recurrent gene is *LAMP2* with two samples showing X-linked rare coding variants. One of the samples is from a male patient with a rare hemizygous missense variant inherited from the mother, while the other sample is from a female patient with heterozygous missense variant also inherited from the mother (Table 3-17). *LAMP2* belongs to the membrane glycoprotein family and constitutes a significant fraction of the total lysosomal membrane glycoproteins [399]. Mutations in this gene have been linked to Danon disease, an X-linked vacuolar cardiomyopathy and myopathy (OMIM 300257)[400].

Other screening studies of *LAMP2* have found mutations in patients with cardiomyopathies [401, 402]. The mouse knockout mouse model did not show gross heart defects, but showed an accumulation of autophagic material in striated myocytes as the primary cause of the cardiomyopathies [403].

Table 3-14 Number of trios with rare coding variants in the ToF replication cohort, classified based on the model of inheritance.

Gene	Autosomal recessive		X-linked
	Homozygous	Compound	
COL18A1		1	
CTBP2		1	
DCHS1		1	
LAMP2			2
MAML1		1	
PCDH15		1	
PCSK5	3		
PLEC		2	
RAI1		1	
ROR2		1	
TCF3		1	

Table 3-15 List of rare coding compound variants in gene. The trio genotypes are represented by 0:homozygous references, 1: heterozygous, 2: homozygous non-reference where the genotype order corresponds to child, mother and father, respectively. VEP: Variant Effect Predictor [170]. 1KG MAF is the minor allele frequency from the 1000 genome project.

Sample ID	**SC_RCTOF5364247**	**SC_RCTOF5364472**	**SC_RCTOF5363671**
Gender	Male	Female	Female
Chromosome	9	9	9
Position	78790207	78790207	78790207
Reference	C	C	C
Alternative	CGAATA	CGAATA	CGAATA
Gene	*PCSK5*	*PCSK5*	*PCSK5*
VEP prediciton	Frameshift	Frameshift	Frameshift
1KG MAF	0	0	0
Trio genotypes	2/1/1	2/1/1	2/1/1
Inherited from	Both parents	Both parents	Both parents

Table 3-16 List of rare coding compound variants in the *PLEC* gene. All variants are missense and predicted by PolyPhen [171] to have a probably damaging (PRD), a possibly damaging (PSD), or a benign (BEN) effect on protein function.

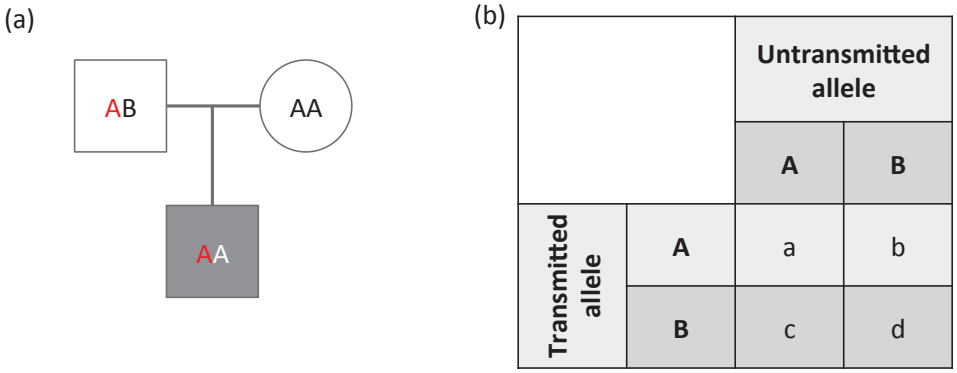
Sample ID	SC_RCTOF5364334		SC_RCTOF5394511			
Gender	Female		Female			
Chromosome	8	8	8	8	8	8
Position	144996830	145003613	144992962	144997315	144998052	144998495
Reference	C	C	G	T	T	G
Alternative	T	T	A	C	A	A
Gene	*PLEC*	*PLEC*	*PLEC*	*PLEC*	*PLEC*	*PLEC*
VEP predication	Missense	Missense	Missense	Missense	Missense	Missense
PolyPhen	PSD (0.856)	Unknown	BEN (0.005)	PSD (0.917)	Unknown	Unknown
1KG MAF	0.004604	0	0.000460	0.00046	0.000921	0.001151
Trio genotypes	1/0/1	1/1/0	1/1/0	1/0/1	1/1/0	1/1/0
Inherited from	Father	Mother	Mother	Father	Mother	Mother

Table 3-17 **List of rare coding compound variants in *LAMP2* gene.**

Sample ID	SC_RCTOF5394505	SC_RCTOF5364097
Gender	Female	Male
Chromosome	X	X
Position	119581776	119581776
Reference	C	C
Alternative	T	T
Gene	*LAMP2*	*LAMP2*
VEP predication	Missense	Missense
PolyPhen	PRD (1)	PRD (1)
1KG MAF	0.003223	0.003223
Trio genotypes	1/1/0	2/1/0
Inherited from	Mother	Mother

3.3.2.6 Transmission disequilibrium test (replication cohort)

Transmission Disequilibrium Tests comprise a group of family-based association tests based on the observed transmissions from parents to affected offspring [404]. The main idea behind a TDT is the ability to detect the distortion in transmission of alleles from a heterozygous parent to an affected offspring (Figure 3-15). The Mendelian analyses above assume complete penetrance and so will not detect inherited variants with incomplete penetrance, but over-transmission of such variants may be picked up by the TDT test.



$$(c) \qquad T = \frac{(b - c)^2}{b + c} \sim X^2$$

Figure 3-15 Original TDT diagram and test statistic. (a) Allele A (in red) transmitted from heterozygous parent to affected offspring. (b) A 2 by 2 table to count all heterozygous parents for the two transmitted alleles and the other two non- transmitted alleles. (c) *T* is McNemar's statistic test and has a chi-square distribution with 1 degree of freedom, provided the sample size of heterozygous parents is sufficiently large. For a smaller number of parents, an exact binomial test can be used [405].

Most, if not all, of the analyses performed in this dissertation are based on the premise that rare coding variants cause CHD including ToF. Without modification, applying the original TDT test on rare coding variants would be underpowered because of the low frequency of these variants (<1% minor allele frequency). To overcome this issue, I modified the TDT test to accept rare coding variants after collapsing their counts per gene in order to increase the power of the test. Once this was done, I generated a 2 by 2 table to calculate *T* of the McNemar's test (Figure 3-15-C) [405]**.** Finally, I obtained a *P* value for each *T* test to decide if a given gene exhibits distorted allele transmission more than expected or not. The *P* values were generated assuming the *T* test has a chi-square distribution with 1 degree of freedom [404, 406].

To create the 2 by 2 table of transmitted and non-transmitted alleles, I consider a child's variant only if it is heterozygous in at least one parent. However, there are many genotype combinations that need to be addressed systematically (Table 3-18). For example, when considering an autosomal chromosome, there are

three possible genotypes: homozygous reference, heterozygous, and homozygous non-reference, which are denoted as 0, 1, and 2 respectively. Because each trio is composed of three members (child, mother and father), there are 27 possible genotype combinations (Table 3-18). Only 13 out of 27 genotype combinations are accepted as TDT informative genotypes and they contribute to the final 2 by 2 table of transmitted and un-transmitted allele counts. The remaining genotype combinations were excluded because they are either not compatible with Mendelian inheritance laws or are non-informative (e.g. when both parents carry homozygous non-reference alleles).

Table 3-18 List of 27 possible genotype combinations in a trio family (homozygous reference, heterozygous, and homozygous non-reference and denotes 0, 1, and 2 respectively). When the status of a genotype combination is non-informative or not compatible with Mendelian laws (the latter is labeled as inheritance error) no rules are applied. However, when a genotype combination is informative (green cells), I add 1 or 2 (under rules) to either transmitted allele or non-transmitted allele counts which both are going to be used in the *T* test.

Genotypes			**Rules**		**Status**
Child	Mother	Father	Add to transmitted alleles count	Add to non-transmitted alleles count	
0	0	0			Non-informative
0	0	1		1	TDT
0	0	2			Inheritance error
0	1	0		1	TDT
0	1	1		2	TDT
0	1	2			Inheritance error
0	2	0			Inheritance error
0	2	1			Inheritance error
0	2	2			Inheritance error
1	0	0			Inheritance error
1	0	1	1		TDT
1	0	2	1	1	TDT
1	1	0	1		TDT
1	1	1	1	1	TDT
1	1	2	1	2	TDT
1	2	0	1	1	TDT
1	2	1	1	2	TDT
1	2	2			Inheritance error
2	0	0			Inheritance error
2	0	1			Inheritance error
2	0	2			Inheritance error
2	1	0			Inheritance error
2	1	1	2	0	TDT
2	1	2	2	1	TDT
2	2	0			Inheritance error
2	2	1	2	1	TDT
2	2	2			Non-informative

Before running the modified TDT test, I made a separate count for each variant class (e.g. frameshift, missense, stop gained, etc.). Since very few silent (or

synonymous) variants are expected to have a sizable effect on the phenotype, I used the transmission of silent variants as an addition control for the TDT tests for both loss-of-function and functional variants with the aim of identifying any technical biases associated with a given gene.

Of the 122 genes selected for the replication study, only one gene, *ARHGAP35*, shows nominally significant over-transmission of rare missense alleles from heterozygous parents to affected offspring (Table 3-19). The modified TDT test reported five rare missense alleles in the *ARHGAP35* gene in the parents (Table 3-20). All of them have been transmitted to the affected children. The rare silent variants in *ARHGAP35* on the other hand did do not show any signs of distorted transmission (six rare silent alleles transmitted and five non-transmitted). However, the difference between missense and silent variants counts are not significant (P= 0.1186, Fisher's Exact test). Given the number of genes tested, the nominal significance of *ARHGAP35* would not survive correction for multiple testing.

ARHGAP35, also known as *GRLF1*, is thought to repress transcription of the glucocorticoid receptor in response to glucocorticoids [407]. This gene was selected in the replication study because I detected one validated *de novo* loss of function in the primary cohort (Table 3-5). The mouse knockout model usually dies within 2 days of birth and does not survive beyond 3 weeks with abnormalities seen in the retina and in the development of the brain and nervous system [408]. Beckerle *et al*. showed how *ARHGAP35* inactivate RhoA, a member of the molecular switches called Rho family GTPases, in response to integrin-mediated adhesion and argued that this inhibition enhances spreading and migration by regulating cell protrusion and polarity [409]. More recently, Kshitiz *et al*. [410] showed how *ARHGAP35* shaped the development of cardiac stem cells, inducing them to become the building blocks for either blood vessels or heart muscle by acting in RhoA-dependent and -independent fashion. These recent findings make *ARHGAP35* an interesting candidate for ToF and CHD in general.

Table 3-19 Transmitted and non-transmitted alleles of rare coding variants in the *ARHGAP35* gene. TDT test were calculated as a McNemar's test (see Figure 3-15).

Gene	Variant class	Transmitted AB	Non transmitted AB	TDT test	*P* Value
ARHGAP35	Functional (missense)	5	0	5.00000	0.02535
ARHGAP35	Silent (synonymous)	6	5	0.09091	0.76302

Table 3-20 List of rare coding missense variants detected in the *ARHGAP35* gene and the genotypes in each trio (child, mother, father). Genotypes are homozygous reference (0) or heterozygous (1).

Chromo.	Position	Reference allele	Alternative allele	Variant class	Genotypes		
					Child	Mother	Father
19	47424846	C	G	Missense	1	1	0
19	47504580	G	A	Missense	1	0	1
19	47422911	C	T	Missense	1	1	0
19	47424531	T	A	Missense	1	0	1
19	47491295	G	A	Missense	1	0	1

Based on the TDT findings in the replication cohort with only 122 genes, I did not perform similar analysis on the primary cohort samples (~20,000 genes), since achieving significant *P* values is not likely after correcting for multiple testing.

3.3.3 Digenic inheritance analysis

I wanted to explore the possibility of digenic inheritance in ToF samples based on two observations. First, there is a well-known example of digenic inheritance with a cardiac phenotype, the long QT syndrome. Patients with long QT syndrome are predisposed to cardiac arrhythmias and sudden death [411]. As with CHD in general, long QT syndrome exhibit locus heterogeneity and variable expressivity but several studies showed a statistically significant digenic inheritance in multiple genes (e.g. *KCNQ1*/*KCNE1* and *SCN51*/*KCNE1*)[412-414]. Secondly, the recurrent *de novo* variants I found in *NOTCH1* and its ligand *JAG1*, although they did not occur in the same patient, they pointed towards the possibility of mutation overload in the same pathway, which I consider in the next section.

To explore this direction, I started by looking for rare coding variants in gene pairs. Because there are ~20,000 genes in the exome data, the search space for gene-pairs is very large (1.9×10^8 unique gene pairs). Even when all possible gene pairs are calculated, the lack of biological evidence to support most of these gene-gene interactions makes it difficult to interpret the results. To overcome this issue, Schaffer has suggested using protein-protein interactions (PPI) to limit the number of possible gene-pairs [351]. I used a list of 68,085 binary PPI integrated from a number of sources by Ni *et al.* [273]. For each pair of genes in the PPI list, I tested two conditions: (i) both genes should include rare, functional, coding variants, and (ii) variant-pairs in affected children are included only if the two variants are inherited from different parents (i.e. similar to the compound heterozygous concept). Rare functional variants are defined as variants with minor allele frequency < 1% in the 1000 genomes project dataset or 2,175 healthy parents from the DDD project, which fall in coding regions or splice sites, and are not synonymous.

This analysis was performed on samples from the primary and the replication cohorts separately. In the primary cohort (n=29 trios), I detected four gene pairs under the DI model that appear in at least two or more trios (Table 3-21). These gene pairs include *TTN*, *OBSCN* and *NEB* genes, which all are giant sarcomeric proteins of striated muscles: titin (*TTN*), nebulette, a member of the nebulin subfamily (*NEB*), and obscurin (*OBSCN*). Mutations in these genes have been linked to cardiomyopathies [415] but the size of these genes is very large and thus it is not unexpected to see an accumulation of rare coding variants in these genes.

Table 3-21 List of interacting gene pairs that carry rare coding variants inherited from one parent in the primary ToF cohort (29 trios). The list below only includes gene pairs that appear in at least two samples.

Gene A	Gene B	Number of trios
MYH2	*OBSCN*	2
GPR98	*MKI67*	2
TTN	*NEB*	2
TTN	*OBSCN*	3

These four gene-pairs are distributed across 8 trios (Table 3-22).

Table 3-22 Breakdown of digenic variant counts per sample in the primary ToF cohort (29 trios)

Sample ID	Gene pairs				Total per sample
	GPR98/MKI67	*MYH2/OBSCN*	*TTN/NEB*	*TTN/OBSCN*	
TOF5136028		1		1	2
TOF5135944				1	1
TOF5135947	1				1
TOF5135980			1		1
TOF5135989		1			1
TOF5135998			1		1
TOF5136004				1	1
TOF5136019	1				1
Total per gene pair	2	2	2	3	9

To test if these findings are statistically significant, I considered 1,080 trios from the Deciphering Developmental Disorders project (DDD) as controls. After performing the same DI analysis on 1,080 DDD trios, I tested each pair of DI genes for a difference in the number of samples between ToF and DDD trios with Fisher's exact test to generate *P* values (Table 3-23). Although some of the DDD trios have heart phenotypes, these are a small minority and I did not exclude these samples from the controls, which makes this analysis more conservative.

Table 3-23 For each pair of genes found in at least two ToF trios (primary cohort), this table list the number of samples from the Deciphering Developmental Disorders project in a given gene pair.

Gene A	**Gene B**	**Cases (ToF n=29)**		**Controls (DDD n=1080)**		**Fisher's Exact Test**	
		Digenic	**No**	**Digenic**	**No**	***P* value**	**Odds ratio**
MYH2	*OBSCN*	2	27	6	1074	0.0168	13.26
GPR98	*MKI67*	2	27	18	1062	0.0938	4.37
TTN	*NEB*	2	27	72	1008	1	1.04
TTN	*OBSCN*	3	26	138	942	1	0.79

None of the gene pairs that include either *TTN* or *NEB* appear to be significant when compared with DDD trios. This indicates that the large size of these genes

is probably the reason why they frequently appear under the DI model and not necessarily because of a pathogenic association.

Only one DI gene pair, (*MYH2*/*OBSCN*), in the primary ToF cohort showed a significant difference (*P*= 0.016) (Table 3-23 and Table 3-24). *MYH2* encodes myosin heavy chain IIa protein and mutations in this gene have been found to cause an autosomal dominant myopathy (inclusion body myopathy-3) [416]. There are six human skeletal MYH genes present as a cluster on chromosome 17 (*MYH1*, *MYH2*, *MYH3*, *MYH4*, *MYH8* and *MYH13*) but only *MYH3* was found to be expressed in the fetal heart and may be involved in the atrial septal defects [417]. Obscurin on the other hand is a sarcomeric protein composed of adhesion modules and signalling domains and surrounds myofibrils [418] but the role of *OBSCN* in cardiogenesis is not obvious [419]. All variants that appear in this gene pair are missense and are predicted to have damaging effects on protein structure (Table 3-24).

Table 3-24 List of rare coding variants in (*MYH2*/*OBSCN*) DI gene pair. All variants are missense and predicted by PolyPhen [171] to have a probably damaging (PRD) or a possibly damaging (PSD) effect on protein function. The genotypes are represented by (0:homozygous references, 1: heterozygous) where the order corresponds to (child/mother/father) genotypes. VEP: Variant Effect Predictor [170].

Sample ID	TOF5136028		TOF5135989	
Chromosome	17	1	17	1
Position	10438612	228566387	10433181	228461504
dbSNP	.	.	rs143872329	.
Ref	T	G	C	G
Alt	C	A	T	A
Gene	*MYH2*	*OBSCN*	*MYH2*	*OBSCN*
VEP	Missense	Missense	Missense	Missense
PolyPhen	PRD (0.971)	PSD (0.317)	PRD (0.915)	PRD (0.993)
AF_MAX	0.00023	0	0.007136	0.002532
Genotypes	1/1/0	1/0/1	1/0/1	1/1/0
Inherited from	Mother	Father	Father	Mother

Because the DI analysis was performed after I designed the replication study, I was not able to include the (*MYH2 / OBSCN)* gene pair in the replication design. However, in my DI analysis in the replication cohort (209 trios) I identified four recurrent DI candidate gene pairs across 11 trios out of 219 possible gene-pairs

available to the 122 genes selected for the replicating study (Table 3-25 andTable 3-26). These four pairs are *ZFPM2/CTBP2, NCOR2/ESR1, PSEN2/NOTCH2*, and *SPEN/NCOR2*. To investigate if any of these gene-pairs were significantly enriched, I compared the number of trios DI variants in these gene pairs between 209 ToF trios and 1,080 DDD trios. Only two gene pairs, *ZFPM2/ CTBP2* and *NCOR2/ESR1* show *P* values < 0.05 (Fisher's exact test, Table 3-27).

Table 3-25 List of interacting gene pairs that carry rare coding variants inherited from one parent in the replication ToF cohort (209 trios). The list below only includes gene pairs that appear in at least two samples.

Gene A	Gene B	Number of samples
NCOR2	*ESR1*	4
PSEN2	*NOTCH2*	2
SPEN	*NCOR2*	2
ZFPM2	*CTBP2*	3

Table 3-26 Breakdown of digenic variant counts per sample in the replication ToF cohort (209 trios)

Sample Id	Gene pairs				Total per trio
	NCOR2 / ESR1	*PSEN2 / NOTCH2*	*SPEN / NCOR2*	*ZFPM2 / CTBP2*	
SC_RCTOF5363452				1	1
SC_RCTOF5363671			1		1
SC_RCTOF5363674				1	1
SC_RCTOF5364163	1				1
SC_RCTOF5364172	1				1
SC_RCTOF5364214				1	1
SC_RCTOF5364247	1				1
SC_RCTOF5364262		1			1
SC_RCTOF5364430		1			1
SC_RCTOF5364460	1				1
SC_RCTOF5394511			1		1
Total per gene Pair	**4**	**2**	**2**	**3**	**11**

Table 3-27 For each pair of genes found in at least two ToF trios (replication cohort), this table lists the number of samples from the Deciphering Developmental Disorders project in a given gene pair.

Gene A	Gene B	Cases (ToF n=209)		Controls (DDD n=1080)		Fisher's Exact Test	
		Digenic	No	Digenic	No	*P* value	Odds ratio
ZFPM2	*CTBP2*	3	206	1	1079	0.0148	15.71
NCOR2	*ESR1*	4	205	5	1075	0.0433	4.2
PSEN2	*NOTCH2*	2	207	1	1079	0.0701	10.43
SPEN	*NCOR2*	2	207	1	1079	0.0701	10.43

The *ZFPM2/ CTBP2* gene pair was mutated in three ToF trios under DI. Whereas *ZFPM2* carries three different missense variants (all predicted to be damaging by PolyPhen [171]) in each trio, *CTBP2* carries the same rare in-frame insertion in all of them (Table 3-28).

ZFPM2, is a known CHD gene and is also called *FOG2*. It is a zinc finger transcriptional factor that is known to regulate many GATA-target genes including *GATA4* in cardiomyocytes [420]. Heterozygous mutations in this gene have been linked to isolated ToF cases [333] and its knockout mouse model shows a spectrum of ToF's structural heart defects [421, 422].

CTBP2, on the other hand, belongs to the C-terminal binding protein family that is linked to multiple biological processes through its association with numerous transcription factors [423]. This gene was picked up during the design process because it is part of the WNT pathway and also because its knockout mouse model showed aberrant halting of heart morphogenesis at the heart tube stage [423].

Table 3-28 List of rare coding variants in the *CTBP2*/*ZFPM2* DI gene pair. All variants are missense and predicted by PolyPhen [171] to have a probably damaging (PRD), a possibly damaging (PSD), or a benign (BEN) effect on the protein function. The genotypes are represented by (0:homozygous references, 1: heterozygous) where the order corresponds to (child/mother/father) genotypes. VEP: Variant Effect Predictor [170]. 1KG MAF is the minor allele frequency from the 1000 genome project.

Sample ID	**SC_RCTOF5364214**		**SC_RCTOF5363452**		**SC_RCTOF5363674**	
Chromosome	10	8	10	8	10	8
Position	126715159	106431420	126715159	106801092	126715159	106456600
dbSNP	.	rs121908601	.	rs202204708	.	rs202217256
Reference	A	A	A	A	A	G
Alternative allele	AGCCGCAGGCTG GGGCTGCAGG	G	AGCCGCAGGCTG GGGCTGCAGG	G	AGCCGCAGGCTG GGGCTGCAGG	A
Gene	*CTBP2*	*ZFPM2*	*CTBP2*	*ZFPM2*	*CTBP2*	*ZFPM2*
VEP	In-frame insertion	Missense	In-frame insertion	Missense	In-frame insertion	Missense
PolyPhen	NA	PSD (0.572)	NA	PRD (0.987)	NA	PSD (0.456)
1KG MAF	0.004374	0.005525	0.004374	0.001381	0.004374	0.004374
Genotypes	1/0/1	1/1/0	1/1/0	1/0/1	1/0/1	1/1/0
Inherited from	Father	Mother	Father	Mother	Father	Mother

The second gene pair that carries rare coding variants under DI model is (*NCOR2/ESR1*) in four ToF trios. When compared with five trios from the DDD project it results in a marginally significant nominal *P* value of 0.043. The *NCOR2* gene, also known as *SMRT*, encodes a silencing mediator (co-repressor) for retinoid and thyroid hormone receptors [424]. This gene was selected in the replication study because it is part of the NOTCH pathway and its null mouse model died before embryonic day 16.5 owing to a lethal heart defect [425].

The second gene in this pair is *ESR1* gene, which encodes for estrogen receptor. Although the *ESR1* knockout mouse model showed no heart structural defects (only decreased heart weight [426]), *ESR1* was included in the replication study because of its role in the NOTCH pathway (reviewed in [427]) (see gene selection in the replication cohort for details). The interaction between *NCOR1* and *ESR1* has been detected by yeast two-hybrid screen assays [428].

All variants in the *NCOR2/ESR1* pair are rare missense variants. With the exception of one variant (rs139960913) that appears in two trios, all other missense variants appear to be unique to each trio (Table 3-29). *NCOR2* also appears in another DI gene pair (*SPEN/NCOR2*), although when compared with DDD trios the difference was not significant (*P* = 0.07).

Table 3-29 List of rare coding variants in the *NCOR2/ESR1* DI gene pair. All variants are missense and predicted by PolyPhen [171] to have a probably damaging (PRD), a possibly damaging (PSD), or a benign (BEN) effect on the protein function. The genotypes are represented by (0:homozygous references, 1: heterozygous) where the order corresponds to (child/mother/father) genotypes. VEP: Variant Effect Predictor [170].

Sample ID	SC_RCTOF5364247		SC_RCTOF5364163		SC_RCTOF5364172		SC_RCTOF5364460	
Chromosome	6	12	6	12	6	12	6	12
Position	152129063	124819118	152129063	124835148	152130253	124835279	152265443	124817779
dbSNP	rs139960913	.	rs139960913	.	rs201212952	rs200297509	rs77797873	rs61754987
Ref	C	T	C	C	A	G	A	C
Alt	T	C	T	T	G	A	G	T
Gene	*ESR1*	*NCOR2*	*ESR1*	*NCOR2*	*ESR1*	*NCOR2*	*ESR1*	*NCOR2*
VEP	Missense	Missense	Missense	Missense	Missense	Missense	Missense	Missense
PolyPhen	PRD (0.996)	BEN (0.311)	PRD (0.996)	PSD (0.72)	BEN (0.001)	PRD (1)	PRD (0.994)	PSD (0.838)
AF_MAX	0.004834	0	0.004834	0	0.005525	0.001381	0.002302	0.003223
Genotypes	1/0/1	1/1/0	1/0/1	1/1/0	1/0/1	1/1/0	1/1/0	1/0/1
Inherited from	Father	Mother	Father	Mother	Father	Mother	Mother	Father

3.3.4 Pathway-based analysis

The final analysis I performed was to test for a burden of rare coding variants in a set of genes linked by biological pathway. To define these pathways, I downloaded the Kyoto Encyclopedia of Genes and Genomes (KEGG) set, which integrates genomic, chemical and systemic functional information to define 175 different pathways [429].

In this analysis, I examined the burden of rare inherited heterozygous missense variants where rare is defined as minor allele frequency < 1% in the 1000 genomes [155] and in 2,172 healthy parents from the Deciphering Developmental Disorders project (DDD) [260]. For each pathway, I counted the number of samples that carry rare missense variants in at least one or more genes from the same pathway. Then, I used Fisher's exact test to detect if the difference between cases and controls is statistically significant.

I applied this workflow on the 29 ToF trios from the primary cohort and used 1,080 trios from the DDD project as controls (Table 3-30). None of the KEGG pathways show a statistically significant burden of rare missense variants after correcting for multiple testing.

Table 3-30 The results of burden analysis from the 29 ToF trios (primary cohort) when considering all genes in the exome data. None of the KEGG pathways reach a significance threshold after correcting for multiple testing (n=175 pathways, adjusted *P* value =0.00028).
FET: Fisher's exact test p-value (right tail), OR: odds ratio

Pathway	# Of genes in pathway	Number of samples				FET	OR
		Cases		Controls			
		> 1 genes	< 1 genes	> 1 genes	< 1 genes		
KEGG_RENAL_CELL_CARCINOMA	12	6	23	100	980	0.05	2.56
KEGG_JAK_STAT_SIGNALING_PATHWAY	7	6	23	107	973	0.07	2.37
KEGG_LONG_TERM_POTENTIATION	7	5	24	94	986	0.11	2.19
KEGG_HUNTINGTONS_DISEASE	5	4	25	68	1012	0.11	2.38
KEGG_PROSTATE_CANCER	11	5	24	97	983	0.12	2.11

On the other hand, the baits in the replication cohort (n=209 trios) only target 122 genes. By performing the same pathway-analysis, but limited to these 122 genes, I was able to detect a burden of rare missense variants in the Dorsoventral axis formation pathway (P=3.4 × 10^{-4}, Fisher's exact test, right tail) and in prion diseases pathway (P =3.6 × 10^{-4}, Fisher's exact test, right tail) (Table 3-31).

Table 3-31 The results of burden analysis from the 209 ToF trios (replication cohort) when considering 122 genes that belong to 73 KEGG pathways. Only 2 of the KEGG pathways reach a significant threshold after correcting for multiple testing (n=73 pathways that have at least 1 gene among the 122 genes, *P*-value threshold=0.00041). The last two rows show the NOTCH and WNT pathways but both of their *P*-values do not reach a statistically significant level. FET: Fisher's exact test, OR: odds ratio

Pathway	# Of genes considered	Number of samples				FET	OR
		Cases		Controls			
		≥1 genes	< 1 genes	≥ 1 genes	< 1 genes		
KEGG_DORSO_VENTRAL_AXIS_FORMATION	4	41	168	114	966	0.00034	2.06
KEGG_PRION_DISEASES	4	24	185	51	1029	0.00036	2.61
KEGG_NOTCH_SIGNALING_PATHWAY	23	99	110	427	653	0.02149	1.37
KEGG_WNT_SIGNALING_PATHWAY	28	62	147	353	727	0.82521	0.86

Next, I tried to see which genes drive the signal of rare missense variants burden in the dorsoventral axis formation and prion diseases pathways. I found four genes (*NOTCH1*, *TP53*, *DLL1*, and *PTPN11*) that show the highest burden of rare missense (Table 3-32). However, only *NOTCH1* reaches a significant *P* value after correcting for multiple testing and it drives the burden signal in both dorsoventral axis formation and prion diseases pathways.

Table 3-32 List of top genes driving the signal of rare missense variant burden in the NOTCH pathway. RMV: rare missense variants, FET: Fisher's exact test

Gene	Cases		Controls		FET right tail	Odds ratio
	With RMV	Without RMV	With RMV	Without RMV		
NOTCH1	22	187	39	1041	8.8 × 10^{-05}	3.1
TP53	4	205	3	1077	0.01	7.0
DLL1	5	204	6	1074	0.02	4.3
PTPN11	3	206	2	1078	0.03	7.8

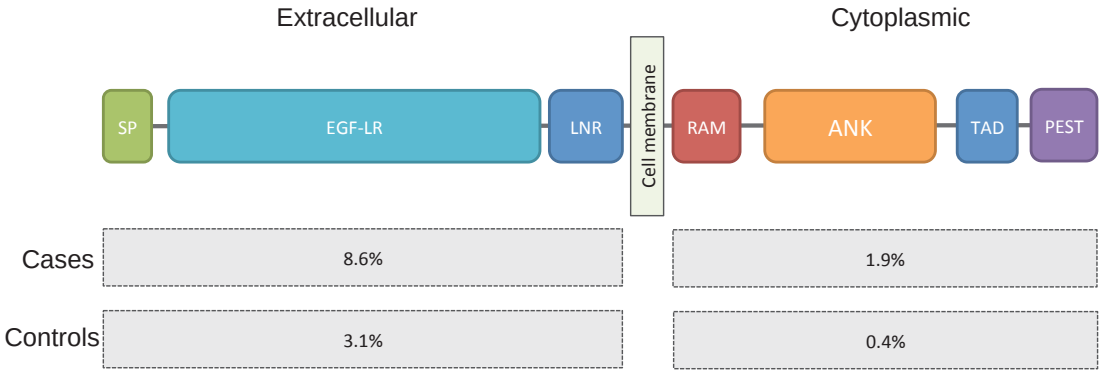


Figure 3-16 Mapping rare missense variants from cases (ToF replication cohort, n=209) and controls (DDD samples, n=1080) to the extracellular or cytoplasmic domains of NOTCH1. The majority of variants are in the extracellular domains where 8.6% of the cases has rare missense variants compared with 3.1% in controls (Fisher's Exact test, *P* value= 0.0007). The number of rare missense variants per domain is listed in Table 3-33 below. SP: signal peptide, EGF-LR: EGF-like repeat, LNR: Lin-Notch repeat, RAM: Rbp-associated molecule, ANK: Ankyrin/CDC10 repeat, TAD: transcription activation domain, PEST: Proline (P), glutamic acid (E), serine (S), and threonine (T) degradation domain.

Table 3-33 Number of samples with inherited rare missense variants in cases (209 ToF trios) and controls (1,080 from DDD) in NOTCH1 domains. The domain boundaries were extracted from Uniport database (protein id: P46531) [430]. LNR: Lin-Notch repeat, ANK: Ankyrin/CDC10 repeat.

Domain	Start	End	Cases (ToF)	Controls (DDD)
EGF-like 4	140	176	1	0
EGF-like 5	178	216	1	0
EGF-like 7	257	293	1	4
EGF-like 13	490	526	0	2
EGF-like 16	603	639	1	3
EGF-like 17	641	676	1	3
EGF-like 18	678	714	0	1
EGF-like 22	829	867	0	10
EGF-like 24	907	943	3	2
EGF-like 25	945	981	1	3
EGF-like 26	983	1019	1	0
EGF-like 27	1021	1057	0	1
EGF-like 28	1059	1095	0	1
EGF-like 33	1267	1305	0	1
EGF-like 34	1307	1346	1	1
EGF-like 35	1348	1384	2	0
EGF-like 36	1387	1426	2	1
LNR 1	1449	1489	1	0
LNR 2	1490	1531	1	1
ANK 2	1960	1990	0	1
ANK 3	1994	2023	0	1
ANK 4	2027	2056	1	0
HIF1AN-binding	2014	2022	0	1

As *NOTCH1* is the most significant gene that I identified in the analyses above, I examined the individual rare missense variants to look for clustering of rare

variants in specific *NOTCH1* domains. The majority of rare missense variants in *NOTCH1* occur in one of the extracellular *NOTCH1* domains (Figure 3-16 and Table 3-33). However, there is no clear domain clustering difference between cases and controls except for the EGF-like 22 domain, which has 10 rare missense variants in DDD control samples and none in the cases. All of EGF-like 22 domain's variants, however, are the same rare missense variant (p.E848K), present in dbSNP (rs35136134). If I omit this SNP, the difference between cases and controls in *NOTCH1* would become statistically more significant (P= 2.9×10^{-6} , Fisher's exact test, right tail).

3.3.5 Summary of candidate genes and gene-pairs

The following table (Table 3-34) summarizes the findings collated from the analyses in this chapter and counts the number of probands (total of 43 candidate genes) under different inheritance scenarios. The most notable gene is *NOTCH1* (n=26 samples) followed by *ARHGAP35* (n=6 trios). Both genes are supported by findings from analyses of both de novo and inherited variants.

Table 3-34 Number of samples with rare coding variants in candidate genes identified in different analyses I performed on the samples from the primary and replication ToF replication studies.
CNV: copy number variant, DN: *de novo*, DI: digenic inheritance analysis, PATH: pathway analysis, R-HOM: Autosomal recessive homozygous, R-COMP: Autosomal recessive compound heterozygous, X: X-lined, TDT: Transmission disequilibrium test. Red cells denote genes with mutation in at least two or more ToF samples.

Gene	Primary cohort (n=29)					Replication cohort (n=209)							Total
	DN	R-HOM	R-COMP	CNV	DI	*DN*	R-HOM	R-COMP	X	TDT	DI	PATH	
ZMYM2	1												1
IKZF1	1												1
TTC18	1												1
MYO7B	1												1
NOTCH1	2					2						22	26
DCHS1	1							1					2
OSBPL10	1												1
TTC18	1												1
FAM178A	1												1
ANKRD11	1												1
ADCY5	1												1
PLCXD1	1												1
ATP5G1	1												1
TPRA1	1												1
FLOT2	1												1
PLCG2	1												1
ARHGAP35	1									5			6
SERAC1	1												1
ITGB4	1												1
PHRF1	1												1
JAG1						2							2
AXIN1						1							1
VEGFA						1							1
PLEC									2				2
COL18A1								1					1
CTBP2								1					1
LAMP2									2				2
MAML1								1					1
PCDH15								2					2
PCSK5							3						3
PLEC								2					2
RAI1								1					1
ROR2								1					1
TCF3								1					1
MYH2/OBSCN					2								2
ZFPM2/CTBP2											3		3
NCOR2/ESR1											4		4
TTN			4										4
OBSCN			2										2
NEB			2										2
HDAC4				1									1
FOXC1				3									3
FOXC2				3									3

3.4 Discussion

Tetralogy of Fallot is the most common form of cyanotic congenital heart defect (~10%) [297]. ToF can occur as part of other syndromes or in isolated non-syndromic forms. Candidate re-sequencing, linkage analysis, CGH arrays, and genome-wide association studies have discovered several novel genes and regions in the past decades. However, the majority of isolated ToF cases remain without definitive genetic causes.

In this chapter, I examined different hypotheses behind the genetic causes of ToF by implementing various, mainly trio-based, analytical tests on the sequence data from 29 isolated ToF trios (exome-sequencing) and later from custom targeted sequencing of 122 genes but in a larger number of samples in a replication study (209 trios).

The quality control (QC) tests in the primary cohort were able to detect a contamination issue in one trio although it had been missed by other quality tests. The various QC reports at the DNA sample processing, sequence data (BAM files) and final called variants (VCF files) proved to be essential steps to remove outlier and contaminated samples before any further downstream analyses. The majority of samples in the replication cohort (n=750) were subjected to whole genome amplification (WGA) prior to sequencing and I did not detect any obvious changes in the quality matrices compared with whole exome sequencing.

The trio study design formed the basis of all analyses discussed in this chapter and not just **detection of *de novo* variants** in the affected children. Although the primary cohort was relatively small (only 29 trios), I was able to detect two *de novo* coding variants (a missense and a single-base deletion of an acceptor splice site) in *NOTCH1*. I also detected one *de novo* missense in another CHD candidate gene, *DCHS1*. Additionally, a novel gene, *ZMYM2*, was found to harbor a *de novo* loss-of-function frameshift. The role of *ZMYM2* in the heart development was

supported by knocking it down in zebrafish using morpholinos by my colleague Sebastian Gerety (appendix A). These functional experiments suggest that zmym2 is essential for normal embryonic heart development, the absence of which causes severe defects leading to death of the embryo. Why do the fish present with such a severe phenotype compared to the patient? While the morpholino injections lead to a loss of correctly spliced mRNA approaching 80-90%, the heterozygous state of our patient, and thus higher level of function protein, could explain the milder phenotype seen, when compared to the zebrafish. Further ongoing work in mouse and zebrafish mutants should clarify these issues.

Collectively, these *de novo* variants explain 13% (4 out of 29 trios) in the primary ToF cohort, which correspond to the predicted proportion of *de novo* variants in CHD cases from a recently published work by Zaidi *et al* [256].

The Mendelian-based analysis of inherited variants using FEVA software identified a few genes with recurrent rare variants under the assumption of complete penetrance. All candidate genes under the recessive model carry compound heterozygous variants in three sarcomeric genes (*TTN*, *NEB* and *OBSCN*). Although these genes have been associated with cardiomyopathies [419], their roles in structural heart defects are not yet confirmed. The large size of these genes is likely to explain why they show up with recurrent rare coding variants.

The burden of rare and *de novo* **Copy number variants (CNVs)** detected by array CGH and SNP arrays are now a well-known cause in 5-10% of isolated ToF cases [340, 341]. Using the read-depth of exome data, CoNVex software was able to detect two *de novo* duplication events, one of which overlaps with *HDAC4* and three inherited small duplications that overlap with *FOXC1* and *FOXC2*. However, they need to be validated using alternative methods first (e.g. custom designed array or multiplex ligation-dependent probe amplification, MLPA). I did not try to call CNVs in the replication cohort since it covers 122 genes only and the CNV boundaries, if any, would be difficult to ascertain. Moreover, most samples were

subject to whole genome amplification, which is known to make calling CNVs robustly in other assays more difficult.

The primary dataset on 29 trios was followed by a **replication study** in 209 trios with isolated ToF. The main goal of this study was to confirm if some of the candidate genes with *de novo* variants might be recurrent in a larger number of isolated ToF samples. Additionally, I wanted to investigate other hypotheses derived from candidate genes published using different methods (GWAS, linkage, animal models, etc.). I selected 122 genes as part of custom designed SureSelect baits from Agilent (USA) for sequencing using an NGS platform (HiSeq, Illumina).

The replication study design based on the number of the genes and the number of sequenced samples would be expected, under the null hypothesis to detect 1.3 *de novo* missense variants and 0.1 loss of function variants and I was able to **identify six *de novo* variants** (half of them are putative loss of function). This suggests an overall enrichment of *de novo* variants of likely functional impact in the selected genes. None of the genes that were selected based on the presence of validated *de novo* coding variants in the primary cohort appeared again in the replication study except for *NOTCH1*. This puts an upper limit on the proportion of ToF that *de novo* variants in these other genes might explain. The replication study shows 1.6% of ToF samples can be attributed to *de novo* coding variants in *NOTCH1* (4 out of 238 trio samples, three are loss of function). This shows a strong over-representation of loss of function variants in the *NOTCH1* gene (P=9.4 ×10^{-8}) given its length and the rate of mutation. Additionally, two *de novo* variants were detected in the *JAG1* gene, a *NOTCH1* ligand, but it did not reach genome-wide significance (P=0.00017), which increases the percentage of isolated ToF cases that can be attributed to *de novo* coding variants in *NOTCH1* or its ligand to 2.5% .

Although I was not able to detect recurrent *de novo* coding variants in other strong candidate genes such as *ZMYM2*, *VEGFA* and *AXIN1*, their biological functions and knockout animal models strongly support their involvement in the heart development and suggest them as novel candidate genes in isolate ToF.

Because of the well-known extreme locus heterogeneity in CHD [431], a larger cohort of isolated ToF trios will be needed to detect additional recurrent *de novo* variants in these genes.

Under **Mendelian inheritance models**, I was able to use the FEVA software to detect three recurrent genes. Three trios carry the same rare frameshift in *PCSK* gene under autosomal recessive homozygous model where all parents are heterozygous. However, because this is an indels, these variant are likely to be false positive due to mapping errors. The second gene was *PLEC* with recurrent compound heterozygous variants, but it is not unexpected for such a large gene. This is similar to what I have already observed in the primary ToF cohort for other large genes (*TTN*, *NEB* and *OBSCN*). However, the rule of *PLEC* gene rare coding variants in congenital heart defects cannot be excluded without further genetic evidence or functional experiments. The third gene was *LAMP2* where I detected rare coding variants under the X-linked model assuming a skewed inactivation of the mother X chromosome. Albeit interesting, this possibility cannot be confirmed without further analysis of the polymorphic androgen receptor (CAG)n repeat region, located on the X chromosome (Xq11-q12) to confirm paternal or maternal X-chromosome skewed inactivation [432] .

To test other variants under a more relaxed scenario of incomplete penetrance, I implemented a modified version of the **transmission disequilibrium test (TDT).** The goal of this analysis was to detect any distortion in the transmission of rare coding variant alleles from heterozygous healthy parents to their affected offspring. Unlike the original TDT, I selected rare functional variants only and collapsed their counts per gene to increase the power of the test. This test detected a distorted transmission of rare missense variants in *ARHGAP35,* a gene recently shown to play a critical role in the development of cardiac stem cells via RhoA-dependent and -independent mechanisms, *in* five trios (~2.4% of the replication cohort). The transmission of rare silent variants in *ARHGAP35* was not distorted like the missense variants but the difference was not statistically significant either. This is most likely because of the small number of variants detected in *ARHGAP35*. Based on these results, the modified version of the TDT

test looks like a promising tool to examine variants with incomplete penetrance. However, a larger sample size is likely to increase the power of this test and make the results statistically more significant. *ARHGAP35* was also suggested as a ToF candidate gene based on the results from the independent *de novo* analysis in the primary cohort where one child has a confirmed *de novo* stop gain variant.

The **Digenic Inheritance (DI) analysis** helped me to explore the area between monogenic and polygenic models, which is rarely considered in CHD genetic literature. The goal of my DI analysis was to detect rare coding variants in gene pairs supported by known protein-protein interactions as long as each variant is inherited from a different parent (similar to the concept of compound heterozygous inheritance but in two genes instead of one). Under the DI model, I identified one nominally significant gene pair from the primary ToF cohort and two nominally significant gene-pairs from the replication cohort. These gene pairs are *MYH2/OBSCN, ZFPM2/CTBP2,* and *NCOR2/ESR1*, all of which are statistically enriched for rare missense variants in ToF samples when compared with 1,080 trios from the Deciphering Developmental Disorders project (DDD). To the best of my knowledge, this is the first systematic DI analysis of genes in any congenital heart defect study. The function and the context in which these gene pairs operate suggest a plausible biological relevance for CHD, especially *NCOR2* and *ZFPM2*. I observed these gene-pairs in 6% in the primary cohort (2 out of 29 in *MYH2/OBSCB*) and in 3% of the ToF replication study (7 out of 209 in *ZFPM2/CTBP2* and *NCRO2/ESR1* gene pairs)

However, a larger sample size is needed to increase the power of any future DI-based analysis. This is especially true for heterogenic disorders such as CHD where hundreds of candidate genes are expected to be involved in the disease. More importantly, functional experiments, either *in vitro* such as cellular assays or *in vivo* (e.g. animal models) are required to confirm the causality of variants under the DI model.

Finally, the **pathway analysis** was more successful in the replication cohort than in the primary cohort. This is probably due to the small number of samples and

large number of genes in the primary cohort. This analysis picked up two pathways: the dorsoventral axis formation and prion diseases pathways. Both of them include the *NOTCH1* gene, which I found to be the main gene driving the signal of rare missense burden in both pathways. *NOTCH1* carries rare inherited missense variants in 22 cases based on this analysis and another four novel rare variants detected by an independent *de novo* analysis (22 out of 238 trios or ~9.2% of all ToF cases).

Although *NOTCH1* is already a well-known CHD gene, its mutations are usually associated with left ventricular outflow tract abnormalities such as aortic valve stenosis, coarctation of the aorta and hypoplastic left heart syndrome [361, 433] more than with ToF cases. My analysis analysis has delineated its contribution to the isolated ToF cases in more detail under different inheritance models including Mendelian, *de novo*, digenic and pathway-based burden. The contribution of each rare missense in *NOTCH1* needs further investigation by means of functional experiments (e.g. luciferase assays, modeling in animals), which are not usually provided for published mutations. These studies would help to determine how the effect of these mutations varies between cases and controls and help us to understand how different mutations cause left or right side structural defects in the human heart.

The analyses described in this chapter also detected other genes with recurrent rare variants under incomplete penetrance in novel genes such as *ARHGAP35* and the *ZFPM2*/*CTBP2* gene-pair under a digenic model. These scenarios represent a partial explanation for part of isolated ToF cases but certainly needs to be confirmed by further genetic evidence and/or functional experiments.

The trio study design has proved to be very informative and a successful design. This design is amenable to many analytical approaches in order to test different hypotheses of the causes of diseases that range from monogenic to polygenic models. A larger sample size of isolated ToF trios will likely prove a productive approach to improving our understanding of the underlying genetic pathogenesis of isolated ToF.