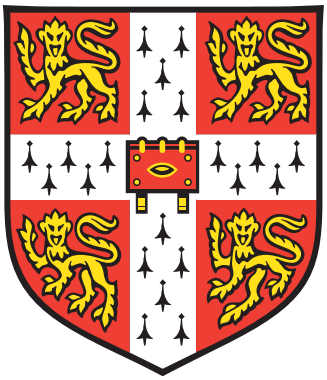
Integrated approaches to elucidate the genetic architecture of congenital heart defects



Saeed Al Turki
Wellcome Trust Sanger Institute
Fitzwilliam College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
September 2013

To Hend, Lma, Leen and Sultan

Declaration

I hereby declare that my dissertation contains material that has not been submitted for a degree or diploma or any other qualification at any other university. This thesis describes my own work and does not include the work that has been done in collaboration, except when specifically indicated in the text.

Saeed Al Turki
26 September 2013

Publications

Publications arising from work associated with this thesis:

- Raffan, E., L. A. Hurst, **S. A. Turki**, G. Carpenter, C. Scott, A. Daly, A. Coffey, S. Bhaskar, E. Howard, N. Khan, H. Kingston, A. Palotie, D. B. Savage, M. O'Driscoll, C. Smith, S. O'Rahilly, I. Barroso and R. K. Semple (2011). "Early Diagnosis of Werner's Syndrome Using Exome-Wide Sequencing in a Single, Atypical Patient." Front Endocrinol (Lausanne) 2: 8.

- Barwick, K. E.*, J. Wright*, **S. Al-Turki***, M. M. McEntagart, A. Nair, B. Chioza, A. Al-Memar, H. Modarres, M. M. Reilly, K. J. Dick, A. M. Ruggiero, R. D. Blakely, M. E. Hurles and A. H. Crosby (2012). "Defective presynaptic choline transport underlies hereditary motor neuropathy." Am J Hum Genet 91(6): 1103-1107.

- Olbrich, H., M. Schmidts, C. Werner, A. Onoufriadis, N. T. Loges, J. Raidt, N. F. Banki, A. Shoemark, T. Burgoyne, **S. Al Turki**, M. E. Hurles, G. Kohler, J. Schroeder, G. Nurnberg, P. Nurnberg, E. M. Chung, R. Reinhardt, J. K. Marthin, K. G. Nielsen, H. M. Mitchison and H. Omran (2012). "Recessive *HYDIN* Mutations Cause Primary Ciliary Dyskinesia without Randomization of Left-Right Body Asymmetry." Am J Hum Genet 91(4): 672-684.

- Schmidts, M., V. Frank, T. Eisenberger, **S. Al Turki**, A. A. Bizet, D. Antony, S. Rix, C. Decker, N. Bachmann, M. Bald, T. Vinke, B. Toenshoff, N. Di Donato, T. Neuhann, J. L. Hartley, E. R. Maher, R. Bogdanovic, A. Peco-Antic, C. Mache, M. E. Hurles, I. Joksic, M. Guc-Scekic, J. Dobricic, M. Brankovic-Magic, H. J. Bolz, G. J. Pazour, P. L. Beales, P. J. Scambler, S. Saunier, H. M. Mitchison and C. Bergmann (2013). "Combined NGS approaches identify mutations in the intraflagellar transport gene IFT140 in skeletal ciliopathies with early progressive kidney Disease." Hum Mutat 34(5): 714-724.

- Gaurav V Harlalka, Anna Lehman, Barry Chioza, Emma L Baple, Reza Maroofian, Harold Cross, Ajith Sreekantan-Nair, David A Priestman, **Saeed Al-Turki**, Meriel E McEntagart, Christos Proukakis, Louise Royle, Radoslaw P Kozak, Laila Bastaki, Michael Patton, Karin Wagner, Roselyn Coblentz, Joy Price, Michelle Mezei, Kamilla Schlade-Bartusiak, Frances M Platt, Matthew E Hurles, Andrew H Crosby (2013). " Mutations in *B4GALNT1* (GM2 synthase) underlie a new disorder of ganglioside biosynthesis". Brain. 2013 Dec; 136(Pt 12):3618-24

- Emma L Baple, Reza Maroofian, Barry A Chioza, Maryam Izadi, Harold E Cross, **Saeed Al-Turki**, Katy Barwick, Anna Skrzypiec, Robert Pawlak, Karin Wagner, Roselyn Coblentz, Tala Zainy, Michael A Patton, Sahar Mansour, Phillip Rich, Britta Qualmann, Matt E Hurles, Michael M Kessels, Andrew H Crosby (2013). "Mutations in *KPTN* encoding kaptin are

associated with autosomal recessive developmental delay with macrocephaly". Am J Hum Genet (94), Issue 1, 87-94

Manuscripts under revision

- D.T. Houniet, T. J. Rahman, **S. Al Turki**, M.E. Hurles, Y. Xu, J. Goodship, B. Keavney, M. Santibanez Koref (2013). "Using population data for assessing next generation sequencing performance". (Bioinformatics)

- **Saeed Al Turki***, Ashok K. Manickaraj*, Catherine L. Mercer*, Sebastian Gerety*, Marc-Phillip Hitz, Sarah Lindsay, Lisa C.A. D'Alessandro, G. Jawahar Swaminathan, Jamie Bentham, Anne-Karin Arndt, Jeroen Breckpot, Jacoba Low, Bernard Thienpont, Hashim Abdul-Khaliq, Christine Harnack, Kirstin Hoff, Hans-Heiner Kramer, Stephan Schubert, Reiner Siebert, Okan Toka, Catherine Cosgrove, Hugh Watkins, Anneke M. Lucassen, Ita M. O'Kelly, Anthony P. Salmon, Frances A Bu'Lock, Javier Granados-Riveron, Kerry Setchfield, Chris Thornborough, J David Brook, Barbara Mulder, Sabine Klaassen, Shoumo Bhattacharya, Koen Devriendt, David F. FitzPatrick, UK10K, David I. Wilson, Seema Mital, Matthew E. Hurles (2013). "Rare variants in *NR2F2* cause congenital heart defects in humans"

Manuscripts in preparation

- **Saeed Al Turki***, Reghan Foley, Sebahattin Cirak, Francesco Muntoni, Matthew Hurles (2013). "FEVA: toolkit for interactive and automated variant prioritisation in family-based exome and genome sequencing projects"

- Katherine J Dick, Emma Baple, **Saeed Al-Turki**, Vijaya Ramachandran, Susan Holder, Matt Hurles, Meriel McEntagart, Andrew H Crosby (2013). "Novel compound heterozygous *WDR62* gene mutations associated with microlissencephaly"

*Join first authors

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Matthew Hurles, for his valuable and constructive suggestions during the planning and development of this research work whilst allowing me the room to work in my own way. His willingness to give his time so generously, insightfulness and critical thinking have kept this project on track. One simply could not wish for a better or friendlier supervisor.

I would like express my gratitude to my other advisor, Dr. Richard Durbin, for welcoming me in his group for my first rotation project, debugging my first Perl script line-by-line and for mentoring and guiding this project through the years. Special thanks to Dr. Inês Barroso and the people in the metabolic disease group for their help during my second rotation project. I also wish to thank my thesis committee: Dr. Lucy Raymond of the University of Cambridge and Dr. Carl Anderson of the Wellcome Trust Sanger Institute.

I'd also like to thank all the people that I have got to know during my time at Sanger: Annabel Smith, Christina Hedberg-Delouka, Alex Bateman and Julian Rayner in the post-graduate office. Sanger's army of pipeline developers past and present, including Shane McCarthy, Petr Danecek, Jim Stalker, Thomas Keane and Carol Scott for keeping the exome data coming my way with ease. Nicola Corton and Carol Dunbar for making sure I remember my deadlines.

I have been blessed through my time with great company, both for enthusiasm for science as well as good times. The people in team 29 have been a source of advice and knowledge: Sarah Lindsay for her valuable help with the validation and screening studies, Parthiban Vijayarangakannan for the CNV calling and being an R mastermind, Sebastian Gerety for his help with the functional experiments. I have enjoyed countless hours of thought-provoking discussions with my colleagues Ni Huang, Marc-Phillip Hitz, Dan King and Arthur Wuster.

This thesis would not have been possible without my collaborators, and I would like to thank them all: Catherine Cosgrove, Jamie Bentham and Shoumo Bhattacharya of the The Wellcome Trust Centre for Human Genetics; Seema Mittal, Lisa D'Alessandro and Ashok Manickaraj from the The Hospital for Sick Children (SickKids), Toronto; Darroch Hall, Bernard Keavney and Judith Goodship from the University of Newcastle; Catherine Mercer and David Wilson from the University of Southampton; David F. FitzPatrick from the University of Edinburgh; Miriam Schmidts, Hannah Mitchison and Peter Scambler from University College London; Andrew Crosby from University of Exeter Medical School; and Chirag Patel and Eamonn R. Maher from the University of Birmingham. I would also like to thank the beta testers of the FEVA program for their valuable input and suggestions: Felicity Payne and Margriet van Kogelenberg. Most importantly, I would like to thank the patients and their families who donated their DNA for all the studies that make up this thesis.

I would like to thank the Wellcome Trust for funding the research in this thesis. My PhD studies would not be possible without the generous scholarship from the National Guard Health Affairs and the UK-Saudi Cultural Bureau in London. My sincerest gratitude to the great people at the department of Pathology in King Abdualziz Medical City: Dr. Mohammed Ali, Dr. Abdulaziz Al Ajlan and Dr. Hanna Bamefleh who were instrumental in opening the first door that led to this PhD and for their encouragement and support. To Dr. Abdulaziz Al Swailem at the KACST who offered me my first job and encouraged me to think big. To my friend, Mustafa Abduallah, for all the great times at the JavaTime planning our half-cooked projects: SaudiBio Inc. and Algenat educational website.

Most of all, I am indebted to my family. Gratitudes in Arabic.

لروح أبي حسين التركي ، أعظم إنسان عرفته ، يا أنقى قلب و يا أصدق الخلق. أعرف أنك لو كنت على قيد الحياة لازددت فخرا بي.. إليك أهدي هذا الجهد. نلقاك عند المولى الكريم الرحيم.

لأمي الحبيبه موزة الدايل، لم تدرسي في مدرسة ولكنك علمتيني كيف اكتب، فشكراً لكل الحروف المُنقطة في دفتري الصغير والتي ساعدتني لأن أكتب هذا الدفتر الكبير .. شكرا لحبك وعطائك الخرافي.

لزوجتي الغالية هند، يدي اليمنى وسندي في الغربة. لقد تكفلتي بكل شيء هنا ولولاك لما استطعت اكمال هذه المرحله في حياتي. اعدك بان اعوضك.

لمهجة قلبي ابنائي لمى ولين وسلطان، لكل اللحظات المرحه معكم التي انتزعتني من ضيق الحياة وصخبها إلى عالم البراءة والطفولة .. آسف عن كل يوم لم اقبلكم قبل النوم وعن كل الساعات التي قضيتها بعيدا عنكم . احبكم جدا.. جدا.

لإخواني ياسر وعبدالعزيز وعبداللطيف وأخواتي أمل ومنيرة ونورة .. شكرا لدعمكم ودعواتكم وحبكم. على الود نلتقي قريبا إن شاء الله

السبت ٢٨ سبتمبر ٢٠١٣م
سعيد بن حسين التركي
كامبردج – المملكة المتحدة

Abstract

Congenital heart defects (CHD) are structural anomalies affecting the heart, are found in 1% of the population and arise during early stages of embryo development. Without surgical and medical interventions, most of the severe CHD cases would not survive after the first year of life. The improved health care for CHD patients has increased CHD prevalence significantly, and it has been estimated that the population of adults with CHD is growing ~5% per year. Understanding the causes of CHD would greatly help improve our knowledge of the pathophysiology, family counseling and planning and possibly prevention and treatment in the future.

Several lines of evidence from humans and animal models have supported a substantial genetic component for CHD. However, gene discovery in CHD has been difficult due to the extreme locus heterogeneity and the lack of a distinct genotype–phenotype correlation. Currently, genetic causes are identified in fewer than 20-30% of the cases, most of which are syndromic while the isolated CHD cases remain largely without explanation.

The aim of my thesis was to identify novel or known CHD genes enriched for rare coding genetic variants in isolated CHD cases and learn about the relative performance of different study designs. High-throughput next generation sequencing (NGS) was used to sequence all coding genes (whole exome) coupled with various analytical pipelines and tools to identify candidate genes in different family-based study designs.

Since there is no general consensus on the underlying genetic model of isolated CHD, I developed a suite of software tools to enable different family-based exome analyses of *de novo* and inherited variants (**chapter 2**) and then piloted these tools in several gene discovery projects where the mode of inheritance was already known to identify previously described and novel pathogenic genes, before applying them to an analysis of families with two or more siblings with CHD.

Based on the tools developed in chapter 2, I designed a two-stage study to investigate isolated parent-offspring trios with Tetralogy of Fallot (**chapter 3**). In the first stage, I used whole exome sequence data from 30 trios to identify genes with *de novo* coding variants. This analysis identified six *de novo* loss-of-function and 13 *de novo* missense variants. Only one gene showed recurrent *de novo* mutations in *NOTCH1,* a well known CHD gene that has mostly been associated with left ventricle outflow tract malformations (LVOT). Besides *NOTCH1*, the *de novo* analysis identified several possibly pathogenic novel genes such as *ZMYM2* and *ARHGAP35,* that harbor *de novo* loss-of-function variants (frameshift and stop gain, respectively).

In the second stage of the study, I designed custom baits to capture 122 candidate genes for additional sequencing using NGS in a larger sample size of 250 parent-offspring trios with isolated Tetralogy of Fallot and identified six *de*

novo variants in four genes, half of them are loss-of-function variants. Both of *NOTCH1* and its ligand *JAG1* harbor two additional *de novo* mutations (two stop gains in *NOTCH1* and one missense and a splice donor in *JAG1*). The analysis showed a strongly significant over-representation of *de novo* loss-of-function variants in *NOTCH1* (P=3.8 ×10^{-9}).

Additionally, when compared with 1,080 control trios, *NOTCH1* exhibit significant burden of inherited rare missense variant (minor allele frequency < 1% in 1000 genomes) (Fisher exact test, P= 8.8 × 10^{-05}) in about 10% of the isolated Tetralogy of Fallot patients. I also modified the transmission disequilibrium test (TDT) to detect any distortion of rare coding allele transmission from healthy parent to their affected children. This modified TDT test identified *ARHGAP35* gene, which exhibits an over-transmission of rare missense variants in children (P=0.025). Although, the p value does not reach a genome-wide significant level after correcting for multiple tests, *ARHGAP35* gene has also a *de novo* stop gain variant in one trio from the primary cohort and recently shown to play a role in cardiomyocyte fate which make it an interesting novel ToF candidate gene for future studies.

To assess alternative family-based study design in CHD, I combined the analysis from 13 isolated parent-offspring trios with 112 unrelated index cases of isolated atrioventricular septal defects (AVSD) in **chapter 4**. Initially, I started with a case/control analysis to test the burden of rare missense variants in cases compared with 5,194 ethnically matching controls and identified the gene *NR2F2* (Fisher exact test P=7.7×10^{-07}, odds ratio=54). The *de novo* analysis in the AVSD trios identified two *de novo* missense variants in this gene. *NR2F2* encodes a pleiotropic developmental transcription factor, and decreased dosage of *NR2F2* in mice has been shown to result in abnormal development of atrioventricular septa. The results from luciferase assays show that all coding sequence variants observed in patients significantly alter the activity of *NR2F2* target promoters.

My work has identified both known and novel CHD genes enriched for rare coding variants using next-generation sequencing data. I was able to show how using single or combined family-based study designs can be an effective approach to study the genetic causes of isolated CHD subtypes. Despite the extreme heterogeneity of CHD, combining NGS data with the proper study design has proved to be an effective approach to identify novel and known CHD genes. Future studies with considerably larger sample sizes are required to yield deeper insights into the genetic causes of isolated CHD.

Table of Contents

Nomenclature

Abbreviations

1KG	The 1000 genomes project
AS	Aorta stenosis
ASD	Septal septal defects
AVSD	Atrioventricular septal defects
CHD	Congeintal heart defects
CNV	Copy number variants
CoA	Coarctation of the
DDD	The Deciphering Developmental Disorders project (www.ddduk.org)
DI	Digenic inheritance model
FEVA	The Family-based Exome Variant Analysis suite
FPR	False positive rate
GAPI	The Genome Analysis Production Informatics
GATK	The Genome Analysis Toolkit (variant calling program)
GQ	Genotype quality
HLHS	Hypoplastic left heart syndrome
INDEL	Insertion or deletion variant
LoF	Loss of function variants
LVTO	Left ventricular outflow tract
MAF	Minor allele frequency
NGS	Next Generation Sequencing
NHLBI-ESP	NHLBI GO Exome Sequencing Project (ESP) ~6,500 exomes
PS	Pulmonary stenosis
QC	Quality Control
QD	Quality by depth
QQ	Quantile-Quantile plot
SB	Strand bias
SNV	Single nucleotide variant
SV	Structural variants
TDT	Transmission disequilibrium test
TGA	Transposition of the Great Arteries
ToF	Tetralogy of Fallot
UK10K	A 10,000 UK-based sequencing project www.uk10k.org
UK10K cohort	Twins cohort study of ~4,000 low-depth genome sequencing project part of the UK10K project
UK10K Neuro	Neurodevelopment sample sets part of the UK10K to study schizophrenia, autism and other psychoses with learning disability
VEP	Variant Effect Predictor
VSD	Ventricular septal defects

List of Figures

List of Tables

1 | Introduction

1.1 Congenital Heart Defects

1.1.1 Historical overview

The chronicle of congenital heart defects (CHD) begins thousands of years ago. The earliest written records of CHD are clay tables dating back to BC 4000 in which the Babylonian listed 62 human malformations and their prophetic implications. One these CHD malformations is *ectopia cordis*, a very rare congenital malformation in which the heart is abnormally located either partially or totally outside of the thorax (Figure 1-1-a), was referred to as follows "*when a woman gives birth to an infant that has the heart open and that has no skin over it, the country will suffer from calamities*" [1].

Generally, one can divide the evolution of our understanding of CHD over the last 300 years to four major eras [2]. The first era extended until the early decades of the 20th century (before the 1940s) and primarily consisted of **descriptive efforts of the pathological anatomy** in the heart (Figure 1-1-b and c). These descriptive efforts culminated when Maude Abbott (Figure 1-1-d) at the McGill University published the first atlas of congenital heart defect in 1936, with detailed clinical and anatomical descriptions of 1,000 malformed hearts [3].

The second era was **of clinicophysiology and surgery (1940s to 1970s)**. The era started when Dr John Streider at the Massachusetts General hospital successfully interrupted a ductus for the first time on March 6, 1937. However, he selected a septic patient who died on the fourth postoperative day of severe pulmonary valve infection (bacterial endocarditis). Because of this regrettable event, Dr Streider halted his regular surgical practice [3]. A year later, on the 16th August 1938, Dr Robert Gross was able to ligate the patent arterial duct in a

7-year old patient who recovered from the surgery and become the first successful patient to undergo heart surgery [4]. In the subsequent couple of years, the work of a team at the Johns Hopkins University revolutionized pediatric cardiology using the opposite operation: instead of closing the ducts as Gross did, they created an artificial duct to rescue cyanotic CHD babies (blue babies) [2]. Although this operation is no longer performed routinely, the whole field of vascular bypass surgery grew from the tools and concepts of their work [2].

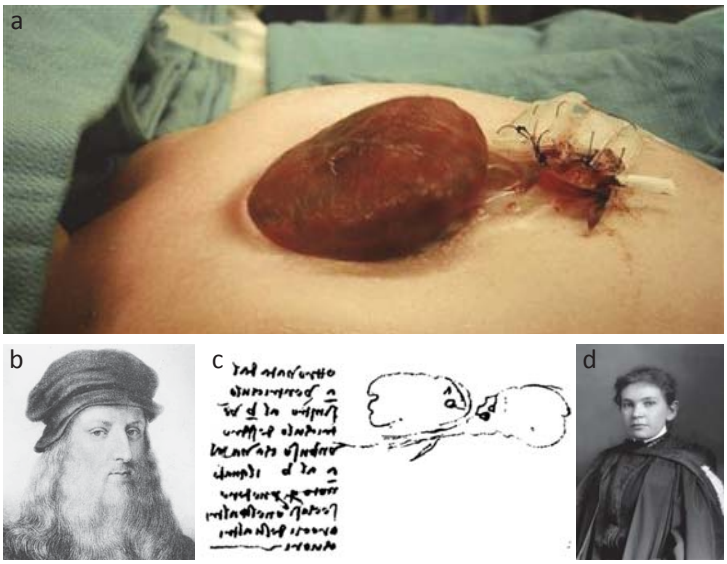


Figure 1-1 (a) A patient with *ectopia cordis*, a malformation mentioned in the Babylonian clay tablets (b,c) An example of anatomical description by Leonardo da Vinci and his drawing of an atrial septal defect in his book, *Quaderni de Anatomia II*. The text read right to left: "*I have found that a, left auricle, to b, right auricle, a perforating channel from a to b, which I not here to see whether this occurs in other auricles of other hearts*" [5, 6]. (d) Maude Abbott in 1869 (image from McCord Museum collection).

The infant era (1970s to 1990s) witnessed the introduction of prostaglandins and the rise of echocardiography [2]. Prostaglandins offered cardiologists a new medical option to keep the ductus open in neonates with various heart defects. The idea was to keep the shunt open to allow the blood to continue circulating until surgery (see fetal circulation section 1.1.9) [7]. The imaging of the heart by ultrasound was another major breakthrough that enabled cardiologists to have a more detailed view of the heart for precise and earlier diagnosis [8].

Researchers in **the current era of cardiac development (1990s and beyond)** have been trying to tackle CHD from different angles. Deep insights into heart development have emerged from multidisciplinary fields such as cellular and molecular biology, human genetics and animal model studies. Methods like linkage, positional cloning, candidate gene sequencing and karyotyping have been used to discover the genetic causes of many syndromic CHD. The results of these studies proved the existence of a clear genetic component in a small proportion of CHD by linking some of the cases to monogenic factors (see CHD genetic causes section 1.1.11.2).

However, epidemiological studies have emphasized a multifactorial (genetic variants interacting with environmental factors) model of CHD causation. Many environmental factors have been found to increase the risk of CHD. One of the most influential studies in this regard is the Baltimore-Washington Infant Study (BWIS) study [9]. This study was a case–control study evaluating genetic and environmental risk factors in live-born infants with CHD in comparison with a control population over a 9-year period. BWIS paints a picture of a wide spectrum of CHD that ranges from monogenic at one end to multifactorial at the other end of the spectrum (see Non-genetic risk factors section 1.1.11.1).

Functional studies have also proven to be an invaluable source of knowledge about heart development. Many ingenious cellular and molecular techniques have been used to dissect the events and processes that take place in heart development. One of these methods is lineage tracing (Figure 1-2) used to follow individual cells at an early stage of the heart development and trace the course of their proliferation and contribution to different heart components. Another method is gene knockdown in zebrafish and mouse knockout models to study how genes and different mutations relate to heart development (see section 1.1.11.2.3).

In the last few years, massively parallel sequencing, also known as next-generation sequencing (NGS), was introduced as a new tool to study different genetic traits in biology and medicine. In this dissertation, I have used NGS to

study some of the non-syndromic CHD that are poorly understood at the genetic level.

This chapter presents an overview of our current understanding of congenital heart defects from the clinical, embryological and genetic perspectives and then describes next generation sequencing methods and their applications.

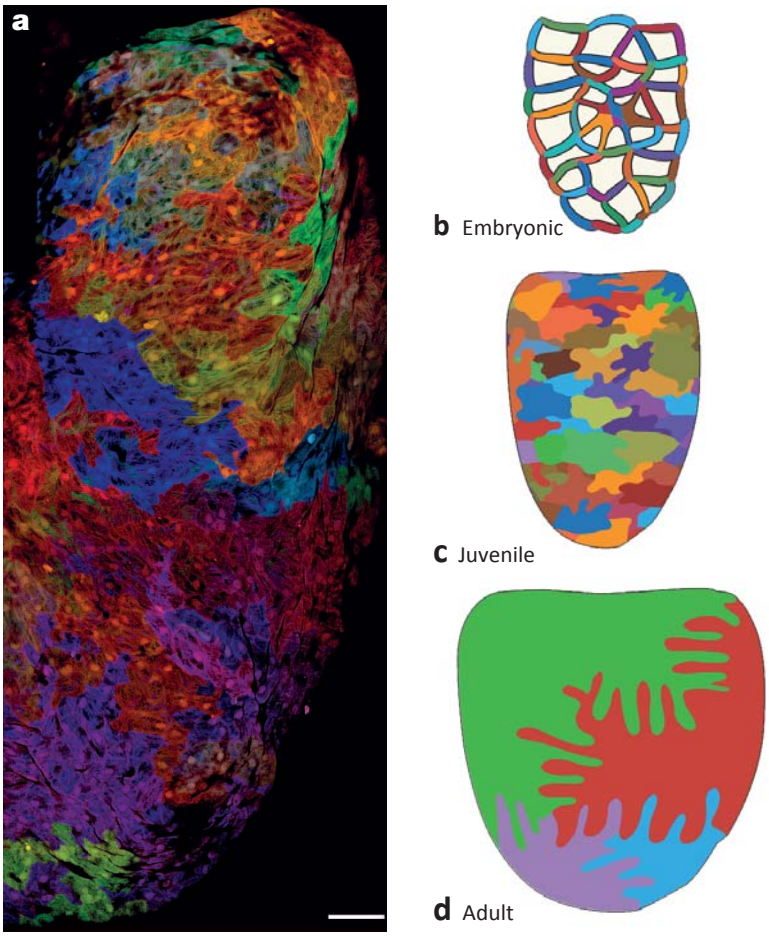


Figure 1-2 Using the Brainbow method [10], a multicolour strategy for following the progeny of numerous individual cells simultaneously, Gupta and Poss [11] show the patterns of cell growth in the zebrafish heart at different stages (a) The embryonic cardiomyocytes that build the juvenile ventricular wall are displayed in clonal patches of variable shapes and sizes [11]. (b) A section through the ventricle of a zebrafish embryo reveals a thin outer wall and an internal meshwork of muscle. Different colours represent different cell lineages. (c) The surface of the juvenile ventricle is an irregular patchwork of multiple lineages. (d) The surface of the adult ventricle is encased by a thick cortical layer that is built by the proliferation of a few founder cells derived from the muscle meshwork (Image 'a' adapted from [11] while the rest were adapted from [12]).

1.1.2 Importance of CHD

Congenital heart defects are considered one of the major health challenges in the 21st century. Collectively, CHD are the most common birth defect with 8-9 new cases in 1000 live births [13] and 1.3 million new cases annually worldwide [14]. Although some heart defects, such as patent ductus arteriosus (PDA), have a minor impact on the patients' life and do not usually require immediate health care, other defects diminish the heart function severely and necessitate intensive medical care and may require multiple surgical interventions.

The prevalence of CHD in adults has been estimated recently at approximately 3000 cases per one million [15] and the size of this population is growing 5% every year [16], in part due to successful surgical intervention during childhood. These figures paint a picture of a major health problem that needs careful planning to accommodate the special medical needs of the CHD patients in the upcoming years.

The impact of a CHD extends beyond the affected child to his family and can lead to catastrophic effects on their psychological and financial welfare. The psychological effect ranges from increased parental stress to severe depression and these complications are usually overlooked [17]. The financial situation of the families may adversely be affected especially in the underdeveloped countries. In one study, a third of the families spend 16% of their monthly limited income on basic medical care and medications to treat chronic heart failure in their CHD child [18].

The causes of the heart defects are largely unknown despite some successes in defining environmental risk factors and genetic causes. The majority of CHD cases remain without definitive diagnosis at the genetic level which hinders medical practitioners from providing optimal health service especially in terms of genetic counselling, family planning, pre-implantation and prenatal diagnosis.

1.1.3 Prevalence of CHD

Since some of the CHD subtypes require an advanced health care infrastructure, planners and policy makers need an accurate estimation of the CHD epidemiological parameters to maintain and expand the medical infrastructure. Towards this end, an extensive body of knowledge has documented the birth prevalence, mortality and complication of CHD (reviewed in [15]). Despite these efforts, most epidemiological studies have been impeded by the variability of CHD definitions, classifications, birth prevalence estimates and survival rates which all led to varied estimates of these parameters.

Epidemiological studies tend to focus on birth prevalence rather than incidence as CHD are congenital defects [15]. The birth prevalence has been estimated as low as four cases up to 50 per 1000 live births [19, 20]. In a country like the USA, the overall estimate of CHD birth prevalence regardless of the subtype is 10 per 1000 live births but if only more severe CHD subtypes are considered, it drops to 1.5 in 1000 live births [19].

On the other hand, the overall prevalence (defined as the number of living patients with the disease in a certain period of time) is more difficult to estimate given the rapid changes in surgical efficacy and survival rates. The estimates in USA and Canada (Quebec) were 3.5 and 4.09 per 1000 adults, respectively, while the prevalence of severe CHD was 0.52 and 0.38 for the same populations [21, 22]. The advances in surgical treatment can change the prevalence as well. The CONCOR registry showed a dramatic improvement of the median age of death, increasing from 37 in 2002 to 57 in 2007 [15, 23]. Currently 96% of newborns with CHD reach an age of 16 because of the improvement in surgical treatment.

1.1.4 Recurrence rate in CHD

The early studies of CHD inheritance in families [24-28], siblings [29] and twins [30] have supported a polygenic or multifactorial model for CHD inheritance.

These studies reported the incidence of CHD in first-degree relatives to be between 1 and 5% [31].

However, the polygenic mode of inheritance was challenged when other studies reported higher recurrence risk (RR) for offspring of patients with CHD. The RR varies considerably among different CHD phenotypes and also varies according to the member of the family who is affected (i.e. sibs, mother or father) (Figure 1-3 and Table 1-1)

For example, when only one child is affected, heterotaxy and TGA show the highest RR (5-6%). Having more than two affected children increases sibling RR up to 10% in ventricular septal defects (VSD) and in hypoplastic left heart syndrome (HLHS). The RR is even more prominent in same sex twins (12 fold) compared to twins with unlike sex [32, 33]. A general observed trend is that hypoplastic left heart syndrome, aortic valve stenosis and coarctation of the aorta (all are obstructive left heart lesions) exhibit higher RR than other CHD phenotypes [34]

Affected parents increase the RR more than having affected sibs but, interestingly, affected mothers result in significantly higher RR compared to affected fathers (2-20% and 1-5% respectively). The reason behind this difference is unknown but epigenetics, imprinting and environmental factors have all been suggested as having a role to play.

The phenotypic concordance of recurrent CHD phenotypes (the same CHD subtype in patients from the same family) is 37% but can be as high as 64% in laterality lesions and 80% in isolated atrioventricular septal defects (AVSD) [35].

In one of the largest population-based studies, Øyen *et al.* examined the familial aggregation of CHD subtypes in a well-defined Danish population that has been annotated in multiple registries. [32]. This study captured all residents of Denmark (~1.7 million) over a 28-year period (1977-2005) and identified ~18,000 individuals with CHD and linked affected individuals with first-, second-

, and third-degree relatives to estimate the contribution of a family history of CHD to an individual's risk of CHD. The authors found the relative risk of recurrence for all types of CHD to be ~3 when a first-degree relative had CHD and diminished when the family history of CHD was in only second- and third-degree relatives which are consistent with the commonly used empirical risks provided to families faced with a potential recurrence of CHD [36]. The same group used the same data to evaluate the general aggregation of dissimilar CHDs in families (by examining all pairwise combinations of discordant 14 CHD phenotypes) and found no evidence that specific combination of the 14 CHD phenotypes aggregated in families [37]. This observation might be explained by the pleiotropic effect of a single gene interacting with external factors (e.g. environmental factors such as pregestational diabetes) and / or interacting with modifier gene(s), which lead to discordant CHDs.

Although Øyen *et al.* have found variable recurrence rate risk for specific CHD (for example the recurrence risk ratio ranged from ~3 in isolated VSD cases to ~80 in heterotaxia), they found that only ~2-4% of heart defect cases in the population were attributed to CHD family history in first-degree relatives. This observation suggests multiple factors, including multiple genetic loci, *de novo* mutations, non-coding factors (e.g. epigenetic), environmental influences, or a combination of these factors are involved in CHD pathogenicity. However, a major limitation of this study, and other similar studies, is that parents with a previous child or other family member with a CHD might be more inclined to opt for prenatal screening and termination of pregnancy if the fetus is affected, which would reduce the observed number of within-family recurrences of CHD and deflate risk ratio estimates accordingly [37].

It has been estimated that 10% of stillbirths exhibit CHD and it is presumed to be a major cause of early fetal loss [14, 38]. RR estimates are thus subject to being biased toward milder forms of CHD since more complex forms of CHD can be incompatible with life. Nonetheless, increased RR in CHD indicates the presence of more familial forms of CHD. The ongoing genetic and molecular studies have indeed confirmed this when rare variants with large effect size have been found

in syndromic and non-syndromic CHD (see section 1.1.11.2 Genetic causes below).

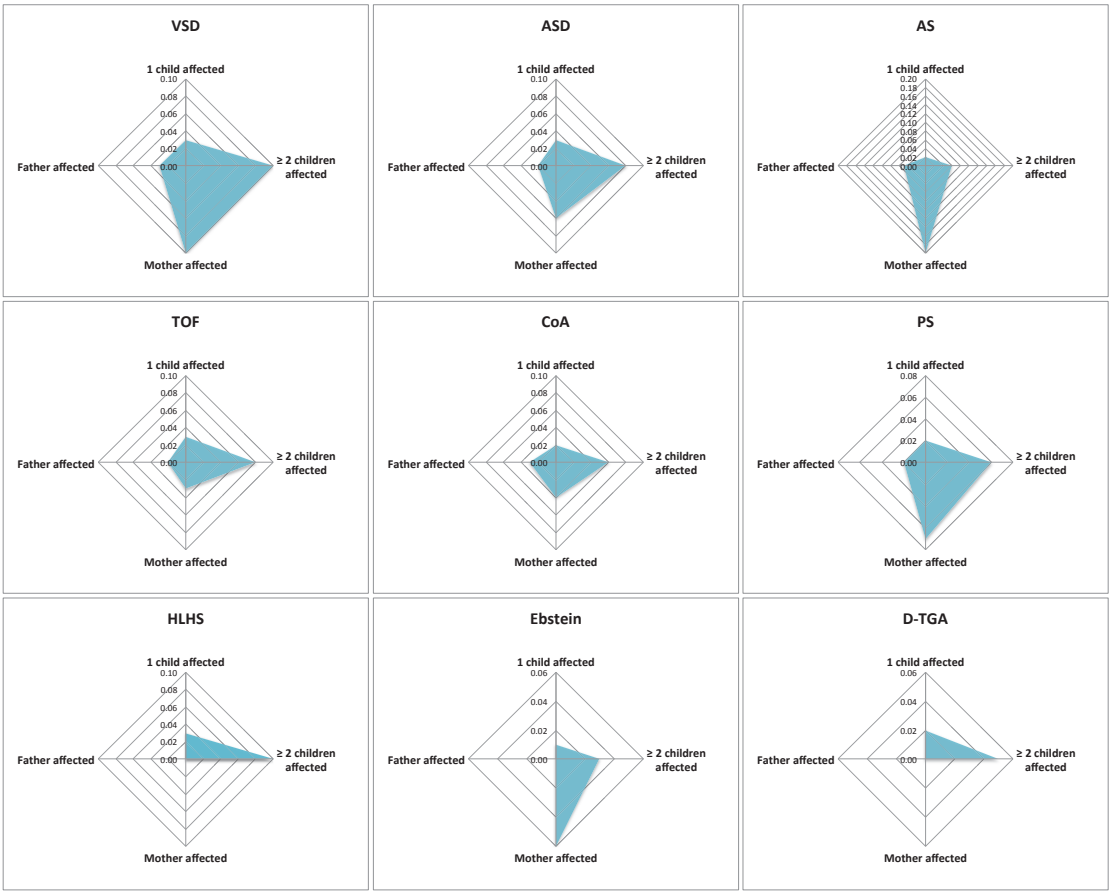


Figure 1-3 Recurrence rate (RR) in selected CHD subtypes. RR value is assigned to 0 when it is not reported [15].

Table 1-1 Recurrence risk (RR) of different CHD subtypes [15, 39]

Cardiac lesion	RR in siblings with unaffected parents		RR in children of affected parents	
	1 child affected	≥ 2 children affected	Mother affected	Father affected
VSD	3%	10%	9-10%	2-3%
ASD	2-3%	8%	6%	1-2%
TOF	2-3%	8%	2-3%	1-2%
CoA	2%	6%	4%	2-3%
AS	2%	6%	12-20%	5%
PS	2%	6%	6-7%	2%
HLHS	3%	10%	nr	nr
AVSD	3-4%	nr	10-14%	1%
PA	1%	3%	nr	nr
TA	1%	3%	nr	nr
TGA	1-2%	5%	nr	nr
L-TGA	5-6%	nr	nr	nr
Ebstein	1%	3%	6%	nr
Heterotaxy	5-6%	nr	nr	nr
Overall	1-6%	3-10%	2-20%	1-5%

ASD = atrial septal defect. AS = aortic stenosis. AVSD = atrioventricular septal defect. CoA = coarctation of the aorta. HLHS = hypoplastic left heart syndrome. L-TGA = congenitally corrected transposition of the great arteries. nr = not reported. PA = pulmonary atresia. PS = pulmonary stenosis. TA = truncus arteriosus. TGA = transposition of the great arteries. TOF = tetralogy of Fallot. VSD = ventricular septal defect.

1.1.5 Clinical presentation and screening for critical cases

About 25% of CHD are considered life threatening and require immediate surgical and palliative intervention in the first year of life [40]. These are usually structural heart defects in which patients are likely to collapse clinically and include transposition of the great arteries, coarctation/interrupted aortic arch, aortic stenosis, pulmonary atresia, and hypoplastic left heart/mitral atresia. It is very important to diagnose these cases as early as possible to provide proper medical care and minimize the life-threating complications.

The early clinical signs of life threating CHD are usually non-specific such as cyanosis (bluish discoloration of the skin), difficulty in breathing and feeding, poor weight gain, and excessive sweating. A cardiovascular examination may reveal abnormal findings such as abnormal heart rate, precordial activity, and heart sounds; pathologic murmurs; and diminished/absent peripheral pulse. The early diagnosis of critical CHD is very important to enhance the survival chances of the affected children. However, it is not always feasible since many critical CHD, especially the ductal-dependent defects, may develop the signs after the initial evaluation and can be easily overlooked [41, 42].

Many newborn screening programs aim to detect pre-symptomatic and critical CHD cases before collapse or death events [43]. Echocardiography is the most sensitive newborn screening method for CHD but it is not cost-effective. A promising alternative is pulse oximetry in the first day, which has been found to improve the early detection of life-threatening CHD [44].

1.1.6 Major health complications of CHD

In infants with untreated complex CHD, most cases of heart failure occur before the end of the first year of life due to volume overload caused by shunts and obstructive lesions of the heart [15]. Heart failure can also occur after surgical treatment such as atrial switch or Fontan procedures in 10–20% of children [45]. Other important late complications of CHD include arrhythmias, endocarditis,

and pulmonary hypertension [46]. Arrhythmias are a leading cause of mortality and morbidity in adults with CHD [47, 48]. Its incidence increases with age and correlates with the severity of CHD [15]. Surgical interventions such as the Fontan procedure can lead to arrhythmias in half of the patients and are thought to arise from trauma to the sinus node and atrial muscle during the surgical procedure [49, 50].

Endocarditis usually arises as a result of the surgical shunts or grafts [51] and its incidence in CHD patients (1.4-11.5 in 1000) is higher than in the normal population (5-7 in 100,000 persons per year). It can lead to serious complications such as valvular regurgitation (30%), cardiac failure (23%), and systemic emboli (20%) [52]. However, earlier surgical treatment and effective use of antibiotics has caused a noticeable decrease in the mortality rate caused by infectious endocarditis to 6-7% [53].

A less common CHD complication is pulmonary hypertension (PH) seen in 4.2-10% of CHD cases [54, 55] which can cause irreversible lung damage if untreated at an early stage. Pulmonary hypertension arises from left-to-right shunting and pulmonary blood volume overload [56]. The high arterial pulmonary blood pressure leads to endothelial dysfunction and increases pulmonary vascular resistance, which leads to central cyanosis (Eisenmenger's syndrome) [57]. The presence of pulmonary hypertension is usually associated with ventricular septal defects [54] and increases the risk of death compared to other CHD patients [58]. Early surgical closure of these shunts helps to decrease the incidence of pulmonary hypertension [15].

1.1.7 CHD classification

Many CHD classifications have been proposed, on the basis of heart structure (anatomical), embryological/developmental, physiological, clinical presentation and/or surgical features. Researchers and clinicians use these classifications to communicate more precisely in different settings. However, there is no

consensus among them on a single CHD classification that is able to capture the complex and multiple facets of congenital heart defects.

One of the most widely used CHD classifications is structure-based (anatomical) and is used in the clinical setting as well as in CHD registries. It also forms the basis of the CHD section in the International Classification of Diseases (ICD 10) [59]. Although a pure anatomical classification is not able to reflect the severity of the diseases, it is very useful when comparing different studies or registries.

A developmental classification of heart defects was used by Leung et al [60] to provide an alternative to the anatomical classifications for obstetricians and ultrasonographers attempting early detection of CHD. This classification is based on detecting deviation from the four-chamber norm and, although it lacks many details captured by the anatomical classification, it is able to provide the correct diagnosis in 97% of CHD cases compared to other methods (post-natal examination, surgery and autopsies) [60]. However, this type of classification is more useful for antenatal diagnosis or to test predictive tools or models of CHD but may change as our understanding of the development of the heart improves [61].

Physiological classifications group CHD by its most significant physiological consequences [62]. For example, cyanotic CHD are characterized by low oxygen levels in arterial blood compared to non-cyanotic heart defects. Such classification is useful for clinical training for simplicity but it overlooks important anatomical features and / or clinical implications [61].

A more useful classification in clinical settings is based on disease severity, suggested by Connelly et al [63] and modified later during the Bethesda conference on congenital heart disease in 2001 [21]. This classification includes three groups – severe, moderate, or simple defects– based on the frequency of an adult CHD patient's visits to a specialized center [15]. This classification was applied to more than seven thousand CHD cases from the PAN registry in Germany (Table 1-2). The majority of cases were mild CHD (~60%) including

small or muscular ventricular septal defects, all types of atrial septal defects, pulmonary stenosis, and patent ductus arteriosus [64].

Table 1-2 Frequency of CHD cases based on clinical severity in 7,245 in newborns (Germany July 2006 to June 2007)

CHD severity	Number of cases	Parentage
Mild CHD	4,372	60.3
Moderate CHD	1,988	27.4
Severe CHD	866	12.0
No classification	19	0.3

Mild CHD include: VSD (small or muscular), ASD (all forms), PDA, PS; moderate CHD include: VSD (others than small or muscular), AVSD, AS, CoA, PAPVC; severe CHD include: UVH (all types), ToF, PA/VSD, PA/IVS, DORV, D-TGA, L-TGA, TAC, IAA, TAPVC, Ebstein's anomaly.
VSD: ventricular septal defects, ASD: atrial septal defects, AVSD: atrioventricular septal defects, AS: aortic stenosis, CoA: coarctation of aorta, PAPVC: partial anomalous pulmonary venous connection, UVH: univentricular heart, ToF: tetralogy of Fallot, PA: pulmonary atresia, PA/IVS: pulmonary atresia with intact ventricular septum, DORV: double outlet right ventricle, D-TGA: dextro-transposition of the great arteries, L-TGA: levo-transposition of the great arteries , TAC: transverse aortic constriction, IAA: Interrupted aortic arch, TAPVC: total anomalous pulmonary venous connection

Although anatomical and clinical classifications are useful, they may obscure developmental relationships in CHD [65]. To address this issue, a pathogenetic classification proposed by Clark [66] was thought be more intuitive when identifying the causes and mechanisms of CHD. Clark's pathogenetic classification includes six mechanisms (Table 1-3). However, a newer version of this classification is needed to reflect the recent insights of heart development research since its last update 17 years ago.

Other classification and coding systems include OPCS 4 (Office for Population Censuses and Surveys) Classification of Surgical Operations and Procedures, Fourth Revision [67] and the European Paediatric Cardiac Code (EPCC) [68] commonly used to code surgical procedures in hospitals in the UK. However, I will adopt the structure-based classification, ICD-10 [59], throughout this dissertation as it is widely adopted and used in clinical practice.

Table 1-3 Clark's Pathogenetic Classification of Congenital Cardiovascular Malformations [66]

Group	CHD
I. Ectomesenchymal tissue migration abnormalities	**Conotruncal septation defects** Increased mitral aortic separation Subarterial, type I ventricular septal defect Double-outlet fight ventricle Tetralogy of Fallot Pulmonary atresia with ventricular septal defect Aorticopulmonary window Truncus arteriosus communis **Abnormal conotruncal cushion position** Transposition of the great arteries (-d) **Pharyngeal arch defects** Interrupted aortic arch type B Double aortic arch Right aortic arch with mirror image branching
II. Abnormal intracardiac blood flow	**Perimembranous ventricular septal defect** **Left heart defects** Bicuspid aortic valve Aortic valve stenosis Coarctation of the aorta Interrupted aortic arch type A Hypoplastic left heart, aortic atresia/mitral atresia **Right heart defects** Bicuspid pulmonary valve Secundum atrial septal defect Pulmonary valve stenosis Pulmonary valve atresia with intact ventricular septum
III. Cell death abnormalities	**Muscular ventricular septal defect** **Ebstein's malformation of the tricuspid valve Group**
IV. Extracellular matrix abnormalities	**Endocardial cushion defects** Ostium primum atrial septal defect Type III, inflow ventricular septal defect Atrioventricular canal defect **Dysplastic pulmonary or aortic valve Group**
V. Abnormal targeted growth	**Anomalous pulmonary venous return** Partial anomalous pulmonary venous return Total anomalous pulmonary venous return and Cor triatriatum
VI. Abnormal situs and looping	**Heterotaxia** **L-loop**

1.1.8 Heart development

The heart is the first organ to develop in the embryo to help circulate nutrients and remove waste. Its development starts as soon as the number of cells reaches a point where diffusion is no longer efficient [69]. Recently, a few techniques have transformed our understanding of how the heart develops. Fate mapping, a method used to determine the cellular derivatives of a cell or population of cells, and lineage analysis in mammalian embryos have documented how different regions in the embryos are involved in cardiac development [70]. This detailed knowledge is likely to improve our understanding of congenital heart defects.

There are four major steps in the development of the heart: formation of cardiac crescent, formation of the heart tube, looping of the heart tube followed by ballooning and finally septation and valve development [70]. These steps result in a four-chambered heart with parallel systemic and pulmonary circulations.

The mature heart consists of different cell types that contribute to structural, biochemical, mechanical and electrical properties of the functional heart. With the help of cell lineage tracing and descriptive embryology of the origins of the heart, researchers have detected **four different populations of cells** that contribute to different parts of the heart [71] (Figure 1-4 and Figure 1-5).

The primary heart field (PHF) forms a cardiac crescent in the most anterior region of the embryo at the second week of human gestation (Figure 1-5-B) [72]. The PHF cells contribute exclusively to the left ventricle and all other parts of the heart, except the outflow tract [73, 74].

The second heart field (SHF) lies medially to the cardiac crescent and then behind the forming heart tube, extending into the mesodermal layer of the pharyngeal arches (Figure 1-5-B). The cells in the SHF contribute exclusively to the outflow tract and all other parts of the heart; except the embryonic left ventricle [71, 73, 75]. It has been suggested that the PHF provides a scaffold upon which cells from SHF migrate into both ends of the heart tube, where they eventually contribute to different cardiac complements [71].

The third source of heart progenitor cells comes from the **cardiac neural crest cells (cNCC)** that migrate as mesenchymal cells into the third and fourth pharyngeal arches and the cardiac outflow (conotruncus) (Figure 1-5-D,E) [76]. The cardiac cNCC cells are necessary for septation of the truncus arteriosus into the aorta and the pulmonary trunk as well as the formation of a part of the ventricular septum [77, 78].

The fourth lineage of cardiac precursor cells is derived from **proepicardium (PE)**, which in turn develops from the coelomic mesothelium that overlay the liver bud (Figure 1-5-E). These cells contribute to the coronary vessels and cardiac connective tissue [71, 79].

The most important events taking place in human cardiac development are listed in (Table 1-4) More details about the development of specific heart structures involved in Tetralogy of Fallot (ToF) and Atrioventricular Septal Defects (AVSD) are discussed in the chapters 3 and 4 of this thesis, respectively.

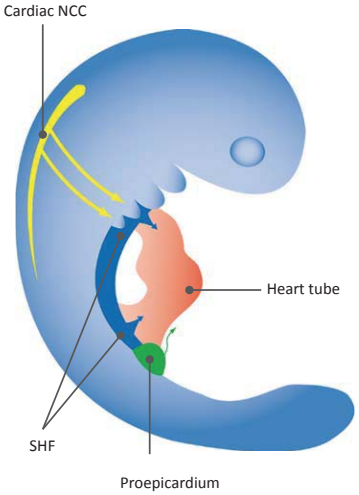


Figure 1-4 Multiple cell lineages contribute to cardiovascular development. A lateral view of embryo at the heart looping stage , around embryonic day (E) 9 in mice, 4 weeks in human, is shown (the image is adapted from [71]).

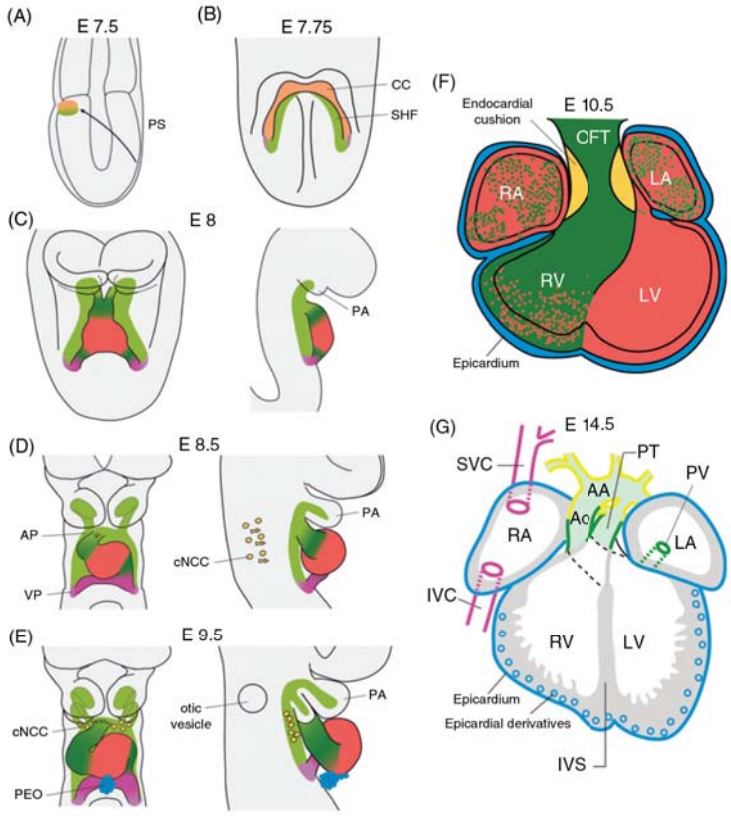


Figure 1-5 (A) Migration of cells anteriorly from the primitive streak (PS). (B) Formation of the cardiac crescent (CC), with the second heart field (SHF) lying medial to it. (C–E) Front (left) and lateral (right) views of the heart tube as it begins to loop with contributions of cardiac neural crest cells (cNCC), which migrate from the pharyngeal arches (PA) to the arterial pole (AP). The proepicardial organ (PEO) forms in the vicinity of the venous pole (VP). (F) The looped heart tube, with the cardiac compartments—OFT, outflow tract; RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle. (G) The mature heart which has undergone septation—IVS, interventricular septum; AA, aortic arch; Ao, aorta; PT, pulmonary trunk; PV, pulmonary vein; SVC, superior caval vein; IVC, inferior caval vein. The primary heart field (PHF) and its myocardial contribution are shown in red, the SHF and its derivatives in dark green (myocardium) and pale green (vascular endothelial cells), cNCC in yellow (vascular smooth muscle of the AA, endocardial cushions), and PEO derivatives in blue. (The image and caption are adapted from [69])

Table 1-4 Stages of human development with corresponding events in cardiac development [80-83]. Carnegie stages are a standardized system of 23 stages used to provide a unified developmental chronology of the vertebrate embryo [83]. DPC: days post coitum.

Carnegie stage	Human DPC	Mouse DPC	Description
CS8	17-19	7	The cardiac crescent forms
CS9	19-21	7.5	The embryo folds, the pericardiac cavity is placed in its final position, gully of myocardium forms, the endocardial plexus forms, cardiac jelly forms
CS10	22-23	8	The heart beats, the endocardial tubes fuse, the mesocardium perforates, looping starts, the ventricle starts ballooning
CS11	23-26	8.5	The atria balloon, the pro-epicardium forms
CS12	26-30	9.5	The septum premium appears, the right venous valve appears, the muscular part of the ventricular septum forms, cells appear in the cardiac jelly, the epicardial growth starts
CS13	28-32	10.5	The atrioventricular-cushions form, the pulmonary vein attaché to the atrium, the left venous valve appears, epicardial mesenchyme appears first in the atrioventricular sulcus
CS14	31-35	11.5	The atrioventricular-cushions approach one another, the outflow ridges become apparent, capillaries form in the epicardial mesenchyme
CS15	35-38	12	The atrioventricular cushions oppose one another, the secondary foramen forms, the distal outflow tract septates the outflow tract ridges reach the primary foramen
CS16	37-42	12.5	The primary atrial septum closes, the outflow tract ridges approach the interventricular septum. The entire heart is covered in epicardium
CS17	42-44	13.5	Secondary atrial septum appears, the sinus node becomes discernable, the left and right atrioventricular connection becomes separate, the proximal outflow tract becomes septated, the semilunar valves develop
CS18	44-48	14.5	Papillary muscles appear, the atrioventricular valves start to form
CS19	48-51	15	The left venous valve fuses with the secondary septum, the mural leaflets of the mitral and tricuspid valve are released
CS21	53-54	16	The main braches of the coronary artery become apparent
CS22	54-56	16.5	The chorda tendinae form
CS23	56-60	17.5	The septal leaflet of the tricuspid valve delaminates

1.1.9 Fetal circulation

The fetal heart blood circulation relies on receiving oxygenated blood from maternal circulation via the umbilical veins (placenta-based) and enters the right atrium of the heart via the inferior vena cava vein. This is facilitated by the presence of two naturally occurring fetal shunts (a connection that allow blood to flow directly from one side of the cardiac circulation to the other), the ductus arteriosus (PDA) and the foramen ovale (PFO) (Figure 1-6). The lungs at this

stage are not developed and have very high pressure that makes the blood divert from the right atrium to the left atrium through PDA and then to the left ventricle and to the rest of the body.

After birth, the first breath increases the O_2 levels in the lungs causing vasodilatation of the lung arteries leading to a sudden drop in the right atrium pressure and an increase in left atrium. This change closes the foramen ovale (becomes fossa ovalis) and similarly, the ductus arteriosus (becomes ligamentum venosum) within 10-15 hours after birth. Postnatally, in 20% to 25%, incomplete fusion leads to the persistence of the flap valve, leaving a PFO opened [84, 85]. Although technically PFO is not a "congenital" defect since it present in all newborns, they are the most common "hole in the heart" among structural heart defects that require catheter intervention [86].

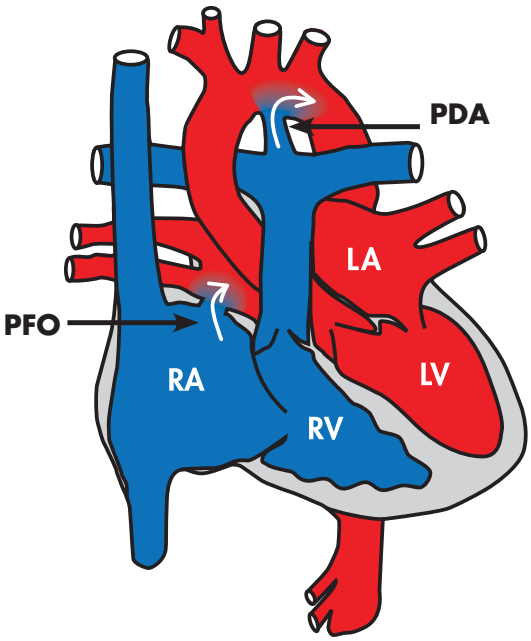


Figure 1-6 The two right-to-left shunts in the fetal circulation, patent ductus arteriosus (PDA) and the patent foramen ovale (PFO) normally closed after birth but may persist longer as symptomatic finding. (Image adapted from Congenital Heart Defects, Simplified (2009) by Ken Heiden [87]).

1.1.10 Anatomical features of CHD subtypes

There are hundreds of subtle anatomical features that have been classified and described in the EPCC and ICD-10 classification systems [59, 68]. This section provides short descriptions of a few selected CHD subtypes because either they are among the most common CHD (e.g. ventricular septal defects, VSD) or are considered severe CHD (e.g. hypoplastic left heart syndrome, HLHS).

Shunts

Shunts are openings between right and left sides of the heart and are considered the most common type of CHD (Figure 1-7-a). The communication can take place between heart chambers, between a chamber and a vessel or between two vessels. They can occur in isolated forms or as part of other severe CHD.

Vessel-vessel shunts

Patent ductus arteriosus (PDA) (Figure 1-7-a, 2) is a naturally occurring communication between the aorta and pulmonary artery. The persistence of PDA is considered the most common form of the CHD but usually does not require surgical intervention when asymptomatic. In cyanotic CHD, the pulmonary circulation entirely depends on the presence of PDA and keeping it open with prostaglandin helps to alleviate the symptoms [88]. Another example is the rare direct communication between the ascending part of the aorta and the pulmonary artery superior to the two semilunar valves called aortopulmonary defect (Figure 1-7-a, 1).

Chamber-vessel shunts

When the upper part of the interatrial septum fails to develop, a sinus venosus atrial septal defect forms and may create a conjunction with the superior vena cava vein (Figure 1-7-a, 3) which is often seen in association with Partial Anomalous Pulmonary Venous Return (PAPVR).

Chamber-chamber shunts

These shunts occur between the ventricles (VSD) or the atrium (ASD) (Figure 1-7-a, 5 to 8). The septum between the two atria contains another naturally occurring shunt in the fetal heart called the patent foramen ovale, PFO, (Figure 1-7-a, 4) and closes immediately after birth (see Fetal circulation section). PFO is a variant of secundum atrial septal defects (ASD) and occurs in the mid portion of the interatrial septum. 20-25% of PFO can persist into adulthood in the absence of other CHD (Figure 1-7-a, 5).

On the other hand, the septum between the two ventricles may rarely have multiple shunts (called "Swiss Cheese VSD"). If it has a single defect at the top of the interventricular septum near the AV annulus it called "membranous VSD" (Figure 1-7-a, 6) or "muscular VSD" otherwise (Figure 1-7-a, 7 and 8).

Atrioventricular septal defects (AVSD)

AVSD is known as endocardial cushion defects or common atrioventricular canal defect and is thought to be caused by the underdevelopment of heart cushions and failure to migrate properly during the development of the heart. ASD and VSD are commonly associated with AVSD along with the abnormal development of the mitral and tricuspid valves (Figure 1-7-b). AVSD classification and further anatomical details are discussed in chapter 4.

Hypoplastic Left Heart Syndrome (HLHS)

This is a cyanotic heart defect caused by severe underdevelopment of the left ventricular, aortic and mitral valves and ascending aorta (Figure 1-7-c). If left untreated, HLHS is responsible for 25 to 40 percent of all neonatal cardiac deaths [89].

Double Outlet Right Ventricle (DORV)

DORV is another cyanotic heart defect characterized by an abnormal origin of both great vessels (aorta and pulmonary arteries) arising either complete or predominantly from the right ventricle. This is usually accompanied by a VSD

that varies in the location and size (subaortic or subpulmonary VSD), which determines the severity of the defect (Figure 1-7-d).

Tetralogy of Fallot (TOF)

TOF is the most common cause of cyanotic complex CHD. It arises by the failure of the interventricular septum to properly attach to the fibrous rings of heart (*anulus fibrosus cordis*) and as a result, causes a misalignment of the infundibulum (the outlet portion of the right ventricular). Four congenital structural defects collectively define TOF: ventricular septal defect, pulmonary stenosis, overriding aorta and hypertrophy of the right ventricular (Figure 1-7-e). TOF is discussed in more details in chapter 3.

Coarctation of the aorta (CoA)

CoA describes a narrowing of the descending aorta, which is typically located at the insertion of the ductus arteriosus just distal to the left subclavian artery (Figure 1-7-f). CoA generally results in left ventricular pressure overload.

Transposition of the great arteries

TGA is another complex cyanotic ventriculoarterial discordant lesion in which the aorta and pulmonary artery reverse their connections to the heart. Normally, the pulmonary artery is located anterior to the aorta and connected to the right ventricle but this is reversed in TGA (Figure 1-7-g). The most common subtype of TGA is the dextro type (referred to as D-TGA) in which the right ventricle is positioned to the right of the left ventricle and the origin of the aorta is anterior and rightward to the origin of the pulmonary artery. A surgical repair is usually required within the first or second week of life.

Ebstein's malformation of the tricuspid valve

This malformation is characterized by downward displacement of the tricuspid posterior and septal leaflets in to the right ventricular. This leads to "atrialization" of the right ventricular as the right atrium becomes enlarged and with a dysfunctional and underdeveloped right ventricular (Figure 1-7-h). The infant's blood circulation may solely depend on the presence of PDA.

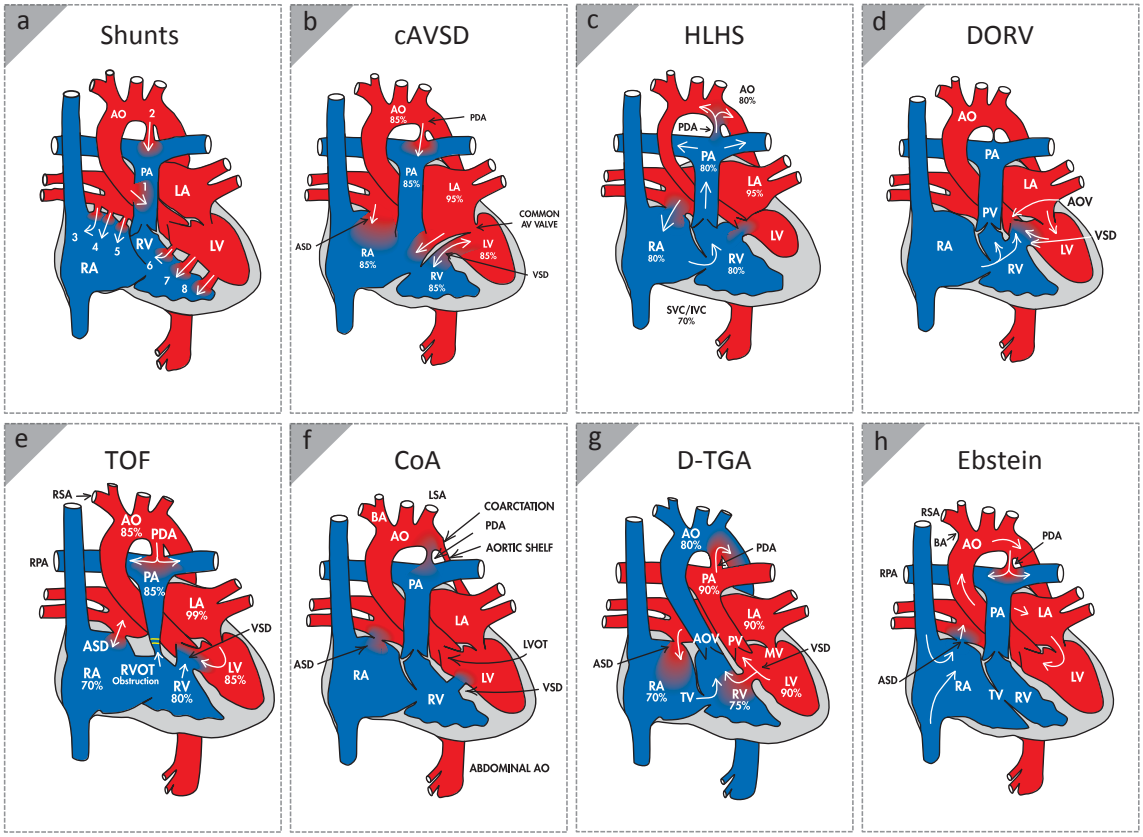


Figure 1-7 Anatomical and physiological features of selected CHD subtypes.
AO: aorta, cAVSD: complete atrioventricular septal defects, D-TGA: dextro-Transposition of the great arteries, DORV:Double Outlet Right Ventricle, HLHS: Hypoplastic Left Heart Syndrome, LA: left atrium, LA: left ventricular, PA: pulmonary artery, PDA: patent ductus arteriosus, RA: right atrium, RPA: right pulmonary artery, RSA: right subclavian artery, RV: right ventricular, TOF: Tetralogy of Fallot.(Images adapted from Congenital Heart Defects, Simplified (2009) by Ken Heiden[87]).

1.1.11 Current understanding of the causes of CHD

1.1.11.1 Non-genetic risk factors

There is a well-established body of epidemiological studies to support several non-genetic CHD risk factors such as maternal rubella; phenylketonuria; exposure to thalidomide, vitamin A, and indomethacin tocolysis [90]. The most influential study in this regard is the Baltimore-Washington Infant Study (BWIS) which was conducted between 1981 and 1989 with a random sample of infants without CHD ascertained from the same birth cohort [9]. This study linked many

environmental factors, different maternal illnesses and certain drugs to the increased risk of CHD.

Pregestational diabetes in particular has been shown to increase the risk of CHD by fivefold with an overrepresentation of transposition of the great arteries, truncus arteriosus, and tricuspid atresia [91]. The exact mechanism is not well understood but several theories have been suggested. One theory suggested high levels of glucose can lead to a disturbance of expression of some master regulatory genes during early embryogenesis [92].

Other factors have been shown to increase the risk of CHD but their impact has varied, and is sometimes contradictory, between different studies. Table 1-5 lists some of the known non-genetic risk factors for any CHD defect when possible; otherwise, I selected the CHD defect associated with highest risk. More details about the association between non-inherited risk factors and specific CHD (TOF and AVSD) are discussed in the third and fourth chapter of this thesis.

Table 1-5 List of the most important non-inherited CHD risk factors.

Risk group	**Factors**	**Heart defects**	**Relative risk**	**Reference**
Maternal illness	Phenylketonuria	Any defects	> 6	[93, 94]
	Pregestational diabetes	AVSD	10.6	[9]
	Febrile illness	Tricuspid atresia	5.1-5.2	[9, 95]
	Influenza	Aortic coarctation	3.8	[96]
	Rubella	Any defects	-	[97]
Maternal drug exposure	Anticonvulsants	Any defects	4.2	[98]
	Ibuprofen	Bicuspid aortic valve	4.1	[99]
	Vitamin A /retinoids	Any defects	-	[100]
Environmental (maternal)	Organic solvents	AVSD	5.6	[9]

1.1.11.2 Genetic causes

To cause a phenotype, the multifactorial polygenic model requires environmental factors to interact with multiple genetic variants each with a relatively small effect size. This model has been widely accepted as the main inheritance model in CHD [28, 39]. However, this view has been challenged by the results of recurrent risk rates in familial CHD, which were found to be higher than what the multifactorial model has predicted. One of the consequences of this discordance is that it has become better appreciated that some proportion of CHD could be explained by monogenic or oligogenic models.

In the past few decades, researchers have utilized various approaches to test different hypotheses and models for genetic causation (Figure 1-8). Classical genetic approaches such as linkage analysis, positional cloning and candidate gene resequencing, that are not generally suitable for dissecting polygenic inheritance have successfully found a genetic cause in 15-20% of CHD cases; most of which have been syndromic CHD [14, 101] (see below).

Only in the last few years, when high-throughput SNP genotyping array (e.g. SNP arrays) were developed, has the contribution of common genetic variants to the polygenic CHD model become amenable to study. Genome-wide association studies have detected a few common variants associated with CHD and this support the continued relevance of the polygenic model (see below).

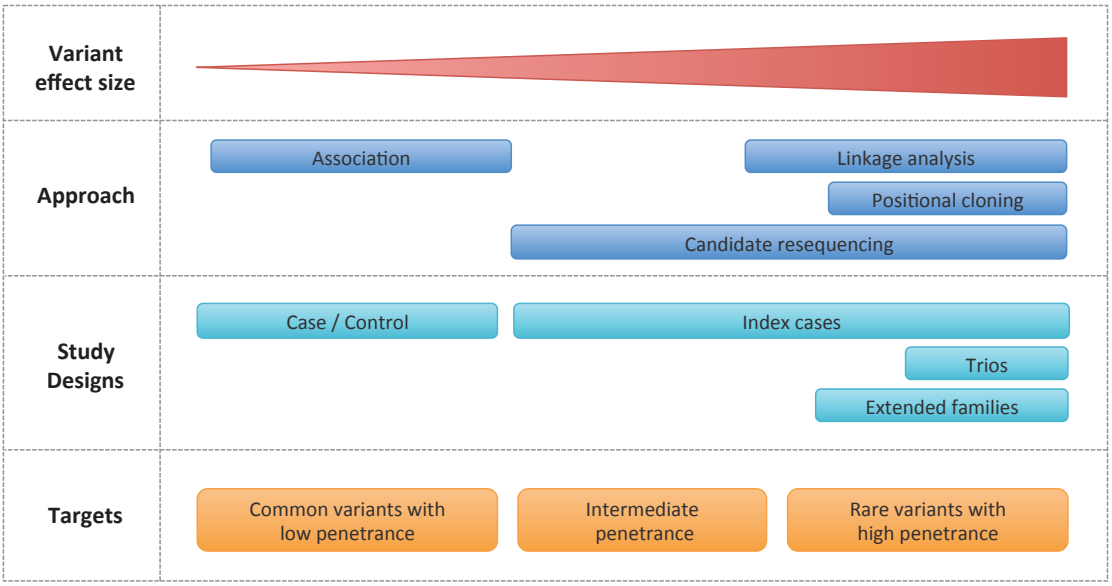


Figure 1-8 Overview of the common DNA-based strategies and methods used to investigate the underlying genetic causes of CHD.

1.1.11.2.1 Syndromic CHD

One or more CHD subtypes can occur as part of a syndrome that also affects systems other than the heart (Table 1-6). The underlying genetic causes of these syndromes can vary from large chromosomal lesions that span multiple genes to single base mutations in a single gene.

About 8-10% of CHD cases are associated with large chromosomal deletions and duplications hundreds of kilobases in length, or greater, that can even involve the whole chromosome as in trisomy 21 (Down syndrome) or monosomy X (Turner syndrome) [102]. It is thought that these large genomic lesions cause CHD when they encompass one or more dosage-sensitive genes where either over- or under-expression leads to a disruption of normal heart development.

For example, the loss of *TBX1* gene in large deletions was found to be responsible for many cardiac phenotypes in Velocardiofacial syndrome [103]. On the other hand, the gain of an extra copy of *RCAN1* gene has been suggested as a partial explanation of CHD subtypes in Down syndrome. *RCAN1* gene is a negative modulator of calcineurin/NFATc signaling pathway that regulates *VEGF-A*

expression, which can be found to cause heart cushion development defects when its expression fluctuates [104, 105].

Another 3-5% of CHD cases are part of different Mendelian syndromes where underlying causes can be attributed to single point mutations, indels and / or microdeletions [20]. For example, Alagille syndrome is an autosomal dominant syndrome defined by the presence of bile duct paucity on liver biopsy and three out of five traits: cholestasis; skeletal, ocular anomalies, characteristic facial features and CHD in 90% of the patients [106]. Coding mutations in *JAG1* gene have been detected in (94%) of the patients [107] while 20p12 deletions were detected in 3-7% [108].

Other syndromes such as Noonan, Holt-Oram, CHARGE and Kabuki have been reported with CHD phenotypes associated with single gene mutations in variable proportions of cases (Table 1-6).

Table 1-6 List of syndromic CHD and the underlying genetic lesions [39, 101]

Causes	Syndrome	Genetic lesion	Cardiac phenotypes	Proportion of CHD
Chromosomal lesions	Edwards	Trisomy 18	VSD, ASD, DORV, TOF, CoA, HLHS	90-100%
	Velocardiofacial	Del 22q11.2	IAA (B), TA, TOF, aortic arch anomalies	75-85%
	Williams	Del 7q11.23	SVAS, PVS, PS, PPS	50-80%
	Patau	Trisomy 13	ASD, VSD, DORV, HLHS, L-TGA, AVSD, TAPVR, dextrocardia, PDA	80%
	Down	Trisomy 21	AVSD, ASD, VSD, TOF	40-50%
	Klinefelter	47,XXY	ASD, PDA, MVP	50%
	Cat eye	Tetrasomy 22p	TAPVR, PAPVR	50%
	Turner	Monosomy X	CoA, AS, HLHS, PAPVR	25-35%
	Pallister-Killan	Tetrasomy 12p	VSD, CoA, PDA, ASD, AS	25%
Microdeletions and Single gene mutations	Hetrotaxy	*ZIC3*	Dextrocardia, L-TGA, AVSD, 90%–100% TAPVR	90-100%
	Alagille	*JAG1, NOTCH1,* del20p12	PPS, TOF, ASD, PS	85-95%
	Noonan	*PTPN11, SOS1, KRAS, RAF1*	PVS, ASD, CoA, HCM	80-90%
	Holt-Oram	*TBX5*	ASD, VSD, AVSD, TOF	80%
	CHARGE	*CHD7, SEMA3E*	ASD, VSD	50-80%
	Char	*TFAP2B*	PDA	60%
	Ellis-van Creveld	*EVC, EVC2*	Primum ASD, common atrium, AVSD	60%
	Smith-Lemli-Opotz	*DHCR7*	AVSD, primum ASD, VSD, PAPVR	45%
	Kabuki	*MLL2*	CoA, ASD, VSD	40%

ASD = atrial septal defect. AS = aortic stenosis. AVSD = atrioventricular septal defect. CoA = coarctation of the aorta. DORV = double outlet right ventricle. HLHS = hypoplastic left heart syndrome. IAA(B) = interrupted aortic arch (type B). L-TGA = congenitally corrected transposition of the great arteries. MVP = mitral valve prolapse. PAPVR = partial anomalous pulmonary venous return. PDA = patent ductus arteriosus. PPS = peripheral pulmonary stenosis. PS = pulmonary stenosis. PVS = pulmonary valve stenosis. SVAS = supravalvular aortic stenosis. TA = truncus arteriosus. TAPVR = total anomalous pulmonary venous return. TOF = tetralogy of Fallot. VSD = ventricular septal defect.

1.1.11.2.2 Non-syndromic CHD

Although isolated non-syndromic CHD are the most prevalent form of CHD, they remain largely without known genetic causes. Linkage analysis and positional cloning have been successfully used in the past few decades to detect some causal genes [14]. The first genes to be reported with autosomal dominant inherited mutations were *NKX2.5* and *GATA4*. Four families with atrial septal defect (ASD) and atrioventricular conduction delay without any apparent non-cardiac features were found to have mutations in *NKX2.5* that were not seen in controls [109]. Similarly, *GATA4* was found to be mutated with novel missense variants in two kindreds with non-syndromic septal defects [110].

Currently, there are 30 genes that have been reported to cause isolated CHD when mutated in humans. Some genes detected with the help of positional cloning include *ZIC3, GATA4, NKX2.5, NKX2.6, MYH6, ACTC1*, and *NOTCH1* while others identified through candidate gene approaches include *TBX1, TBX20, CFC1, CITED2, CRELD1, FOG2, LEFTY2, NODAL, GDF1, FOXH1, TDGF, MYOCD, TLL1, THRAP2* and *ANKRD1*. These genes can be arranged into three classes based on their functions: transcriptional factors, receptors/ligands and structural protein (Table 1-7) Most of the mutations detected were missense variants inherited in an autosomal dominant fashion with variable penetrance.

One major limitation of some classical genetic approaches such as linkage analysis is that it requires large extended families with multiple affected family members. The rarity of such large CHD families limits the use of these approaches and has led researchers to look for alternative methods in their quest to discover the genetic causes of CHD.

Table 1-7 List of genetic models and genes associated with non-syndromic CHD [111]

Model	Gene group/class	Gene / Locus	Cardiac phenotypes
(a) Presumed high-penetrance autosomal dominant mutations	Ligand-receptor	*NOTCH1*	BAV, AS
		CFC1	Heterotaxy, TGA, TOF, TA, AVSD
		LEFTY2	Heterotaxy
		ACVR2B	Heterotaxy
		GDF1	TOF
		ALK2	ASD, TGA, DORV, AVSD
		NODAL	Heterotaxy
		TDGF1	TOF
		JAG1	PS, TOF
	Transcription factor	*GATA4*	ASD, TOF, VSD, HRV, PAPVR
		GATA6	PTA, PS
		NKX2.5	ASD-AV block, TOF, HLHS, CoA, IAA, Heterotaxy, TGA, DORV, VSD, Ebstein
		NKX2.6	PTA
		TBX20	ASD, CoA, VSD, PDA, DCM, MS, HLV, ASD
		CITED2	VSD, ASD
		FOXH1	TOF, CHM
		ZIC3	Heterotaxy, TGA, ASD, PS
		TBX5	ASD, VSD, AVSD
		TBX1	VSD, IAA
		ANKRD1	TAPVR
	Contractile proteins	*MYH11*	PDS, AA
		ACTC1	ASD, VSD
		MYH6	ASD
		MYH7	ASD, Ebstein
		MYBPC3	ASD, VSD
	Miscellaneous	*FLNA*	XMVD
		ELN	SVAS
		TLL1	ASD
		THRAP2	TGA
(b) Common variants with low penetrance	Methylation cycle	*MTHFD1*	TOF, AS
		MTRR	Various
		SLC19A1	Various
		NNMT	Various
		TCN2	Various
	Vasoactive proteins	*NPPA*	Conotruncal defects
		NOS3	Conotruncal defects
	Polypeptide mitogen	*VEGF*	VSD, PTA, IAA, TOF
	Transcription factor	*NFATC1*	VSD
		MSX1	ASD [112]
(c) Somatic mutations	Gap junction protein	*GJA1*	HLHS
	Transcription factors	*NKX2.5*	VSD, ASD, AVSD
		GATA4	VSD, AVSD
		TBX5	ASD, AVSD
		HEY2	AVSD
		HAND1	HLV, HRV
(d) Copy Number Variations (CNVs)	*De novo* and / or inherited gain or loss	1q21.1	TOF, AS, CoA, PA, VSD
		3p25.1	AVSD
		4q22.1	TOF
		5q14.1-q14.3	TOF
		9q34.3	TOF, CoA, HLHS
		19p13.3	TOF

One alternative method is to detect association between CHD phenotypes and specific loci or common variants. For example, by searching for association

between common variants in 23 candidate genes and non-syndromic Tetralogy of Fallot (TOF), Goodship *et al.* found a single variant (rs11066320) in *PTPN11* that increases the risk by 5% [113]. Rare mutations in *PTPN11* are known to cause Noonan syndrome, which includes congenital heart disease, by up regulating Ras/mitogen-activated protein kinase (MAPK) signaling. A few other common variants were found to be associated with the increased risks of certain types of CHD in (Table 1-7, b).

A more powerful approach is to perform genome-wide association studies (GWAS) using SNP arrays. GWAS have been very successful in general; they have found more than 8,500 genome-wide significant associations across more than 350 human complex traits such as Diabetes Mellitus Type 2 and obesity [114]. Unfortunately, this level of success has not been matched thusfar in CHD, except for two published examples [112, 115]. Cordell *et al.* [112] found a moderate signal of association with the risk of ostium secundum atrial septal defect (340 cases) with p-value of ($P = 9.5 \times 10^{-7}$) near the *MSX1* gene. Although this study had a relatively larger number of CHD cases of various types (1,995 in total) and has the power to detect moderate-sized effects; it failed to find a globally strong signal when combining all CHD types. Only after the team analyzed the phenotypes separately, did the signal reached a genome-wide significant level and accounted for 9% of the population-attributable risk of ASD and suggested that genetic associations with CHD may exhibit considerable phenotypic specificity.

Zhibin Hu *et al.* published the second example of GWAS in CHD patients from Han Chinese population [112, 115]. Their multi-stage GWAS study included 4,225 CHD cases and 5,112 controls in total and found two strong signals near *TBX15* and *MAML3* genes.

This modest performance of GWAS in CHD is not unexpected due of the heterogeneity of CHD phenotypes. Large collaborations between national CHD registries and large cohorts of homogeneous clinical CHD cases are expected to improve the discovery rate of associations [14].

Non-Mendelian inheritance mechanisms have also been suggested to explain some isolated CHD. The **somatic mutations** and two-hit hypothesis suggested by Knudson has been widely accepted in tumor neology and skin diseases. Later studies by Reamon-Buettner and Borlak show somatic mutations in *NKX2.5, TBX5, GATA4, HEY2* and *HAND1* from the human heart tissue [116, 117]. However, subsequent work by Draus *et al.* failed to replicate these findings in fresh frozen tissues from 28 septal defect patients. They suggested that the poor DNA quality from the formalin-fixed tissues in the work of Reamon-Buettner and Borlak was the source for these somatic mutations [118]. However, this doesn't eliminate a possible role for somatic mutations in CHD, but their involvement remains to be confirmed by additional larger studies.

Small noncoding microRNAs (miRNAs) have also emerged lately as important players in cardiogenesis [119, 120]. These are short 20 to 26 nucleotides, evolutionary conserved RNAs that usually interact with the 3' untranslated region (UTR) of specific target mRNAs to control their expression. Their involvement in heart development processes, such as cardiac pattering, angiogenesis, and cardiac cell fate decisions have been documented by many studies (reviewed by [119]). The upregulation of four maternal miRNA (miR-19b, miR-22, miR-29c and miR-375) were found to be associated with congenital heart defects in the fetus and thus have been suggested as non-invasive biomarkers for the prenatal detection of fetal CHD [121].

Most recently, ***de novo* variants** of different classes have been shown to contribute to as much as 10-15% of CHD cases. Soemedi *et al.* observed rare *de novo* CNVs in 5% CHD-affected families [122]. Additionally, whole exome sequencing of 362 trios detect recurrent *de novo* mutations (base substitutions and indels) in several genes including *SMAD2* [123]. Although this cohort include both syndromic and isolated CHD, based on the expression of the mutated genes in the developing heart compared to genes mutated in control trios, the authors estimated that in 10% of patients the *de novo* mutations contributed to the CHD.

1.1.11.2.3 Known CHD genes in mouse

In addition to the human genetic approaches described above, studying the effect of knocking out genes in mouse models and how it affects the heart development has identified 300 genes that when homozygously knocked out result in abnormal cardiac development [124]. Additionally, a combination of high-throughput imaging systems (MRI) and ENU mutagenesis workflow has enabled researchers to screen thousands of mice per year and to generate a list of candidate genes for resequencing in humans. Extrapolating from the mouse knockout data, based on the current incomplete coverage of mammalian genes, it has been estimated that the total number of genes that when homozygously knocked out cause CHD in the mice may be 1,500-2,000[124].

1.2 Next generation sequencing (NGS)

Before 2004, the DNA sequencing field was dominated by automated Sanger sequencing, also known as 'capillary sequencing', which has been considered the first generation of sequencing [125]. Capillary sequencing helped to generate the first human genome (2.8Gb with 99% completion and 1 in 100,000 error rate)[126]. Despite its great success, it is considered a low-throughput technology, expensive, and labor-intensive for large-scale projects. A new wave of novel sequencing approaches started in 2005 when the first commercially available massively parallel sequencing platform was released by Roche/454 [127] and the multiplex polony sequencing protocol of George Church's lab [128].

These new waves of high-throughput approaches were labeled "next-generation sequencing", which refers to a combination of advancement in the chemistry, sequencing, signal detection, imaging and computation methods that allow

researchers to generate a vast amount of biological data (DNA- or RNA-based sequencing data) in a short time and at a reasonable cost [129].

Currently, there are several commercially available platforms: Roche/454, Illumina/Solexa, Life/SOLiD, Helicos BioSciences, Polonator instrument and Pacific Biosciences among many others. Each of these platforms adopts various methods to sequence the DNA such as pyrosequencing, reversible terminator, sequencing by ligation. Each has its own advantages and disadvantages in terms of the length of DNA fragments, ease of preparation, error rates, run time and the amount of data they produce per run in Giga-bases. These methods can be grouped into a few categories: (i) microelectrophoretic methods [130], (ii) sequencing by hybridization [131], (iii) real-time observation of single molecules [132, 133] and (iv) cyclic-array sequencing [134] (reviewed by Michael Metzker [135] and Shendure *et al.* in [136]). However, the sequencing itself represents the first few steps in a larger workflow.

1.2.1 A standard NGS workflow

The standard NGS workflow is composed of multiple steps or tasks that can be arranged in two main categories: laboratory-based and computational-based. The laboratory steps include DNA preparation, library quality control and sequencing. The computation-based tasks start with converting raw sequencing signals (e.g. images or electrical changes) to text-based DNA sequence reads, mapping to the genome, calling variants, quality control, filtering, annotation and finally specialized down-stream analysis based on the biological question and the study-design (e.g. trios, case/control) (Figure 1-9).

This workflow is commonly shared between different sequencing platforms [137]. However, I will discuss this workflow with the Illumina/Solexa platform in mind since it is currently the most widely used platform [135] and was the only platform used to sequence the samples in this thesis.

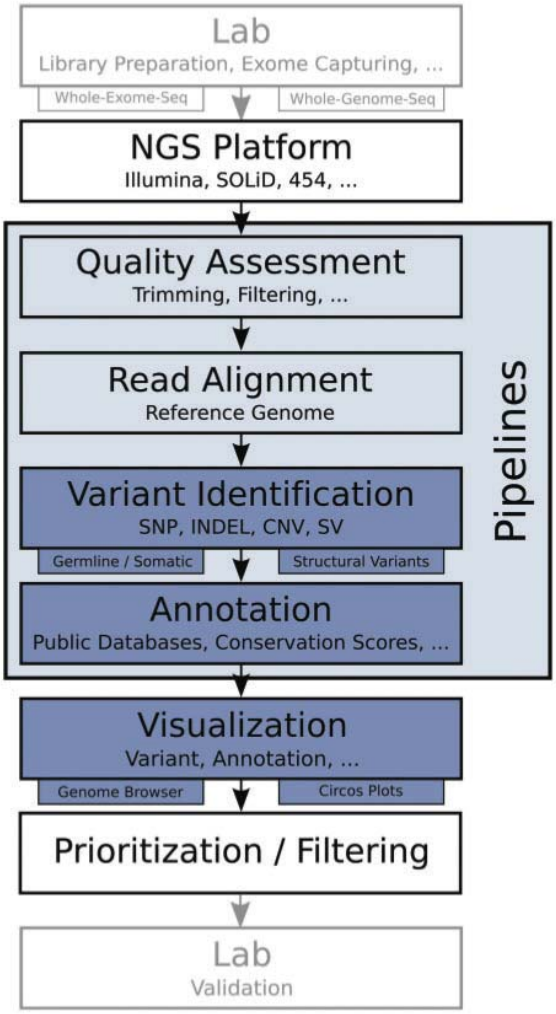


Figure 1-9 Basic workflow for whole-exome and whole-genome sequencing projects. After library preparation, samples are sequenced on a certain platform. The next steps are quality assessment and read alignment against a reference genome, followed by variant identification. Detected mutations are then annotated to infer the biological relevance and results can be displayed using dedicated tools. The found mutations can further be prioritized and filtered, followed by validation of the generated results in the lab. (The image and caption are adapted from [137])

1.2.1.1 Laboratory-based steps

The laboratory based steps start with genomic DNA extracted from blood, saliva or tissue samples. The amount and concentration of DNA required for sequencing depends on the platform and the size of targeted regions (e.g. whole exome or whole genome). For example, for the work described in this thesis, targeted exome sequencing on HiSeq Illumina platform required 2000 ng of DNA. In addition to DNA volume and concentration, an electrophoretic gel is also used

to check for DNA integrity. At the early stages, DNA contamination should be checked rigorously before proceeding any further. One approach to test for possible DNA contamination issues is to genotype a handful of autosomal and sex chromosomal SNPs to match gender and test relatedness.

Library preparation is accomplished by DNA fragmentation using physical (ultrasonic) or chemical approaches [138] into smaller pieces of relatively homogenous length followed by ligation to common adaptor sequences. To empower signal detection during sequencing, clonally clustered amplicons need be generated using *in situ* polonies, emulsion PCR or bridge PCR among others methods [136]. The goal of these methods is to generate multiple copies of a single DNA molecule arranged spatially on a planar substrate or bead surface.

Sequencing specific parts of the genome (e.g. all coding regions as in the whole-exome) requires capturing these regions with predefined baits of various lengths (90-mer in the case of TruSeq Exome Enrichment Kit from Illumina and 120-mer in SureSelect Exome Enrichment Kit form Agilent). To increase the number of samples sequenced per run (8, 16, 24, 48 and 96), some of the exome enrichment protocols add an indexing step to allow samples to be pooled but their data deconvoluted.

Once a library is ready, massively parallel sequencing is based on enzyme-driven biochemistry and imaging-based (SOLiD, Solexa) or voltage-based data acquisition (Ion Torrent) (see Table 1-8 for more details about different platforms).

Table 1-8 Technical specifications of some commercially available Next Generation Sequencing platforms [139, 140]

Platform	**MiSeq**	**Ion Torrent PGM**	**PacBio RS**	**HiSeq 2000**	**SOLiD 5500xl**	**FLX Titanium**
Company	Illumina	Life technologies	Pacific Biosciences	Illumina	Life technologies	Roche / 454
Instrument Cost	$128K	$80K	$695	$645K	$251K	$450K
Amplification method	Bridge PCR	Emulsion PCR	None	Bridge PCR	Emulsion PCR	Emulsion PCR
Sequencing method	Sequencing by synthesis	Sequencing by synthesis (H^+ detection)	Sequencing by synthesis	Sequencing by synthesis	Ligation and two-base coding	Pyrosequencing
Data acquisition	Image-based	Semiconductor-based	Image-based	Image-based	Image-based	Image-based
Sequence yield per run	1.5-2Gb	1Gb (318 chip)	100 Mb	600Gb	155 Gb	0.4 Gb
Sequencing cost per Mb*	$0.07	$1.20	$2-17	$0.04	$0.07	$12.00
Run Time	27 hours	2 hours	2 hours	11 days	8 days	10 hours
Primary errors	Substitution	Indel	Indel	Substitution	A-T bias	Indel
Observed Raw Error Rate	0.8%	1.7%	12.8%	0.3%	≤ 0.1%	1.0%
Read length	Up to 150 bases	~200 bases	Average 1,500 bases	Up to 150 bases	75+35 bases	Up to 700 bases
Paired reads	Yes	Yes	No	Yes	Yes	No
Insert size	Up to 700 bases	Up to 250 bases	Up to 10 kb	Up to 700 bases	NA	NA
Typical DNA requirements	50-1000 ng	100-1000ng	~1 µg	50-1000 ng	NA	NA
* The prices are updated as of 2013 [139, 141]						

1.2.1.2 Computation-based steps

The first computational step starts by converting the raw signals detected by NGS platforms (e.g. the fluorescence in imaging-based systems) to sequence reads, 'base-calling'. This step usually takes place on or next to the sequencing machine in real time. The output is composed of raw sequence reads in addition to the corresponding quality score for each base in a file format called "FASTQ" [142].

Each sequencing platform suffers from different types of error during base-calling [143]. For example, the 454 platform infers the length of homopolymers from the observed fluorescence intensity, which varies and usually leads to

higher error rate with indels (short DNA insertion or deletion variants). The Illumina platform on the other hand has a miscall rate around 1% due to different errors. As the Illumina read sequence length increases, the DNA synthesis process desynchronizes between different copies of DNA templates in the same cluster and base-calling becomes less accurate in later cycles. Because of these errors, reads with an excess of sequence artifacts, base calling errors and adaptor contamination need to be excluded before mapping them to the human genome reference[144].

The remaining high quality reads are then mapped to one of the available human genome references such as the Genome Reference Consortium human build 37 (GRCh37). Many alignment tools have been developed in the last few years to map millions of DNA sequencing reads (reviewed by [145, 146]). The majority of the fast aligners generate auxiliary data structure called indices for the reference sequence, the read sequences or both [145]. Based on the indexing method, these aligners can be arranged into three groups: hash tables-based aligners such as BALT [147] and SSAHA2 [148], suffix trees-based aligners such as BWA [149] and Bowtie [150], and merge sorting-based aligners such as Slider [151] .

BWA was used to align raw sequence reads from all samples discussed in my thesis. BWA generates Sequence Alignment/Map (SAM) files [152], a tab-based format that describes the alignment of reads in rich detail. SAM files include two parts: a header for metadata (optional) and an alignment section. Each line in the alignment section describes one sequence read in details: where it maps on the reference genome, the quality scores at base and read levels, a CIGAR string to record the matching output between the read bases and the reference genome and many other additional pieces of information. A binary version of SAM file format, called BAM, is usually preferred over SAM format to save digital storage space and provide faster operations and queries.

Before calling variants from sequencing reads in BAM files, a few additional quality control steps are usually applied to reduce the false positive rate (FPR). For example, base quality score recalibration attempts to correct the variation in

quality with machine cycle and sequencing context, as implemented in GATK [153, 154]. Once this is done, the quality scores in the BAM files are closer to the actual probabilities of erroneously mismatching with the sequenced genome. Additionally, removing reads with excess mismatches to the reference genome, realignment around common insertion/deletions and discarding duplicate reads originating from a single progenitor template can enhance the FPR. These steps generate BAM files with high quality reads that are ready for variant calling and many of them have been developed as part of the 1000 genome project [155].

Today, there are more than 60 variant callers available (reviewed by [137]). These callers can be arranged into four groups according to the type of DNA variant: (i) germline callers (discussed below), (ii) somatic mutation-calling based on DNA from matched tumor-normal patient samples are an essential part of many cancer genome projects (reviewed by Kim and Speed [156]), (iii) copy number variant callers from NGS (reviewed by Duan *et al.* [157]) , and (iv) structural variants (SV) callers which are designed to call insertions, deletions, inversions, inter- and intra-chromosomal translocations (reviewed by Pabinger *et al.* [137]).

Germline callers include GATK [153, 154], Samtools [152] and they are used to call single nucleotide and short indels. These programs call a variant at a given locus when it is sequence different from the reference genome and then they try to determine its genotype status based on the number of alleles (heterozygous, hemizygous or homozygous non-reference in the case of human DNA). Initially, simple algorithms based on allele counts at each site were used to call a variant or genotype using simple cutoffs. Recently, uncertainty was incorporated in more sophisticated statistical frameworks for variant / genotype calling [143]. Because indels suffer from higher false positive rates, additional Bayesian-based (e.g. Dindel [158]) or pattern-growth based programs (e.g. Pindel [159]) may be used to improve their calling and genotyping (reviewed by Neuman *et al.* [160]).

The germline callers usually output variant and genotype calls in a standardized generic format for storing sequenced variants including single nucleotide, indels,

larger structural variants and annotations called Variant Call Format (VCF) [161]. The VCF format is easily extendable and is able to hold rich details about every variant in single or multi-sample files. VCF can be compressed by up to 20% of its original size to save storage space and also can be indexed (e.g. using Tabix [162]) for fast random access which is essential for most downstream analyses.

The number of variants in VCF files depends on the size of sequenced regions. The numbers can range from four million variants in deep whole genome sequences to about 40-80 thousand variants in whole 50Mb-size exomes. This large number of variants represents a challenge when researchers try to look for genetic causes of disease. Additional filtering and annotation are usually applied to exclude unwanted variants. For example, population allele frequencies from public resources such as the 1000 genomes project [155] or NHLBI GO Exome Sequencing Project (ESP) [163] are useful to exclude common variants (e.g. minor allele frequency > 1%). Comparative genomics provides a base-resolution conservation score (e.g. GERP [164, 165], phastCons [166] or phyloP [167]). These scores are useful when analyzing non-coding variants since most important functional elements of the genome are expected to be more conserved.

Since most high penetrance pathological variants occur in coding regions (i.e. exons) as reported by human genetic mutation database (HGMD) [168], predicting the variant effect on protein structure is an important part of any downstream analysis. SNPeff [169] as well as Variant Effect Predictor (VEP) from Ensembl [170] are two commonly used programs used for this task. More specialized tools are used to predict the damaging effect of missense mutations such as PolyPhen [171], SIFT [172] and Condel [173].

These annotations and filters, along with computation approaches discussed in chapter 2, can help to minimize the search space for plausible casual variants dramatically, by order of magnitudes, down to few tens or hundreds of candidates per sample.

1.2.2 NGS applications

NGS has revolutionized many fields such as microbiology, molecular biology, population genetics, cancer genetics and molecular diagnostic to name a few. Although NGS applications have been extended with greater success to non-human organisms such viral, bacterial, plants, and animals, this section focuses on human-related applications only.

Broadly speaking, NGS applications in humans can be divided into two groups: medical-based and research-based applications (Figure 1-10). There is a thin-line between these two groups as many of the studies or applications start out as a research-based, but once a solid foundation is established, they are usually translated into clinical practice.

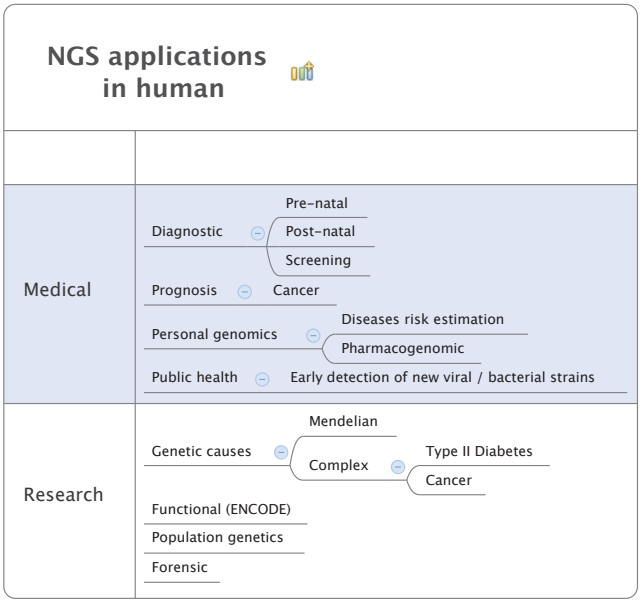


Figure 1-10 Examples of NGS applications in human

Monogenic genetic disorders

This is probably one of the most active research areas where NGS has been demonstrating great success. Ng *et al.*, in 2010 [174] showed for the first time how NGS was able to show that mutations in *DHODH* gene cause Miller syndrome, a recognized autosomal recessive disorder. Since then, the genetic

causes of tens of rare Mendelian disorders have been deciphered under autosomal recessive, dominant inherited, dominant *de novo* and X-linked models (see Table 1-9 for more examples).

Table 1-9 Selected studies using exome and whole genome sequencing for disease gene identification [175]

Sequencing	Inheritance Model	Disease	Putative Loci Identified	Reference
Exome	Autosomal dominant	Familial amyotrophic lateral sclerosis	*VCP*	[176]
		Neonatal diabetes mellitus	*ABCC8*	[177]
		Primary lymphedema	*GJC2*	[178]
		Spinocerebellar ataxia	*TGM6*	[179]
	Autosomal recessive	Carnevale, Malpuech, Michels, and oculoskeletal-abdominal syndromes	*MASP1*	[180]
		Charcot-Marie-Tooth neuropathy	*GJB1*	[181]
		Congenital chloride losing diarrhea	*SLC26A3*	[182]
		FADD deficiency	*FADD*	[183]
		Familial combined hypolipidemia	*ANGPTL3*	[184]
		Fowler syndrome	*FLVCR2*	[185]
		Joubert syndrome 2	*TMEM216*	[186]
		Mental retardation	*TECR*	[187]
		Miller syndrome	*DHODH*	[174]
		Nonsyndromic hearing loss (DFNB82)	*GPSM2*	[188]
		Seckel syndrome	*CEP152*	[189]
	Sporadic	Mental retardation	Several genes	[190]
		Schinzel-Giedion syndrome	*SETBP1*	[191]
	X-linked recessive	Intractable inflammatory bowel disease	*XIAP*	[192]
Genome	Autosomal dominant	Metachondromatosis	*PTPN11*	[193]
	Autosomal recessive	Charcot-Marie-Tooth neuropathy	*SH3TC2*	[194]
		Miller syndrome	*DHODH, DNAH5,* and *KIAA0556*	[195]
		Sitosterolemia	*ABCG5*	[196]

Whole exome sequencing (WES) is the preferred method in most of these studies for its low cost and smaller number of variants compared with whole genome sequencing (WGS). Unlike WGS, where non-coding variants are the dominant variant type, WES targets coding regions of the genome (~1-2%), which enhances interpretability of the variants and can be subjected to further analysis with functional experiments.

Researchers have used a common strategy to find the causal genes in these studies. This strategy usually starts by comparing the WES/WGS variants with public databases such as the 1000 Genomes Project [197, 198], the NHLBI Exome Variant Server[199], International HapMap Project [200], and single-nucleotide polymorphism (SNP) database (dbSNP) [201], as well as internal controls [202]. By focusing on rare variants (typically with minor allele frequency < 1% in controls), this usually excludes most of the variants in WES, down from ~20,000 coding variants to a few hundreds.

The detection of rare coding variants in the same gene in unrelated individuals or families with the same monogenic disorder is usually considered strong evidence to support the causality. However, additional functional studies are usually needed to support the pathogenicity if the candidate mutation appears only in a single-family [202].

To date, more than 180 novel genes have been linked to monogenic disorders using next-generation sequencing where the causal mutations were either occur *de novo* or inherited [202]. Different family designs ranging from unrelated cases, affected sib-pairs and trios have been used to investigate different inheritance models (Table 1-10). Autosomal recessive disorders were over-represented during the first few years (2009-2011) of using NGS platforms to elucidate causes of monogenic disorders. This over-representation was mainly due to the fact that a small number of affected sib-pairs are enough to find the causal homozygous variants. In non-consanguineous families that demonstrate an autosomal recessive inherence pattern, the exome data from one or two sib-pairs were usually enough to find a few compound heterozygous variants to be the cause of the disease (see the example of *DDHD2* gene in Table 1-10). In consanguineous families, 15-20 rare homozygous candidate variants are expected in affected sib pairs [202].

Similarly, autosomal dominant disorders caused by *de novo* mutations are relatively easy to identify using a parent-offspring trio design. This analysis requires exome data from the affected child and both parents and is usually less

complex since few *de novo* variants are present in each sample (for example *EZH2* gene in Table 1-10).

Familial autosomal dominant disorders are more challenging because of a large number of rare heterozygous candidate variants per sample. Sequencing larger numbers of affected samples and / or coupling with linkage analysis in extended families can help to minimize the number of candidate variants. For example, a 2.9Mb linked region detected in a large family (32 affected members with Familial Diarrhea Syndrome) was targeted for sequencing in only 3 affected members. The coupling of linkage analysis and NGS resulted in detecting a rare single heterozygous missense variant in the *GUCY2C* gene.

Table 1-10 Example of gene identification approaches and study designs coupled with NGS to elucidate the genetic cause in some of the published monogenic disorders in the least 2-3 years.

Inheritance model	**Study design**	**Analytical approaches**	**Examples of monogenic disorders**		
			Disorder	**Gene**	**Number of cases/families**
Autosomal recessive	Affected sib-pairs	- Shared homozygous or compound heterozygous in affected sibs and heterozygous in unaffected parents	Complex form of hereditary spastic paraparesis [203]	*DDHD2*	One affected sib-pair
Consanguineous autosomal recessive	Affected sib-pairs	-Shared homozygous variants and heterozygous in unaffected parents - Identical By Decent (IBD) analysis (Autozygosity)	Postaxial polydactyly type A [204]	*ZNF141*	Three affected sibs in one family of a Pakistani origin
X-linked recessive	Affected male child and healthy mother	- Shared variants in affected males and carrier mothers.	Diamond–Blackfan anaemia [205]	*GATA1*	Two affected male children and a carrier healthy mother
Autosomal dominant	Affected parent-child or unrelated index cases	- Co-segregation of heterozygous in affected parent-child. - Variant in the same gene in unrelated families. - NGS coupled with linkage analysis in large families	Familial Diarrhea Syndrome [206]	*GUCY2C*	Captured a 2.9 Mb linked region in 32 members of a large Norwegian family
***De novo* dominant mutations**	Complete trios	- *De novo* variant in child not seen in healthy parents	Weaver syndrome [207]	*EZH2*	Two unrelated parent–child trios

Cancer

Many studies have utilized NGS platforms to detect genes with recurrent somatic mutations in different solid and hematological neoplasms [208], acquired somatic mutations in melanoma [209] , substitution and rearrangement in lung cancer [210, 211] and in breast cancer [212].

Recurrent somatic mutations in *DNMT2*, for example, were detected in 22% of patients with Acute Myeloid Leukemia (AML) [213]. These mutations provide not only a deep insight into the tumor biology but also have a prognostic value. Patients with *DNMT2* mutations were found to have a worsened prognosis when they have a normal cytogenetic profile [213]. Additionally, pilot studies have successfully adapted NGS to monitor the cancer progression by detecting the residual disease following treatment [214, 215]. This was based on sequencing of immunoglobulin VDJ gene rearrangements in lymphoma or lymphoid leukemia for minimal residual of disease (MRD) [214].

Multifactorial disease

Very recently, Morrison *et al..* used low-coverage whole-genome sequencing of 962 cases to study the genetic architecture of a complex trait, levels of high-density lipoprotein cholesterol (HDL-C) [216]. Their results showed 61.8% of the heritability of HDL-C levels could be attributable to common variations. This supported the hypothesis that common variants are likely to represent true polygenic variations with small effects. The use of NGS to find these common variants is expected to play an important role in identifying the biological pathways involved in the complex disease pathophysiology.

Infectious disease

Identifying novel infectious organisms and tracking outbreaks or epidemics of disease requires a fast and thorough response before they become a major health problem. NGS platforms fit the bill perfectly and have proved their

tremendous value in such situations. The 2010 Haitian cholera outbreak was traced to have originated in Bangladesh using NGS [217]. Similarly, the Escherichia coli O104:H4 break in Germany were found to be a Shiga toxin-producing strain [218]. The underlying mechanism behind its virulence was thought to arise by horizontal transfer of a prophage carrying genes for Shiga toxin 2 and other virulence factors [218].

More recently, NGS enabled the discovery of a novel Middle East Respiratory Syndrome (MERS) coronavirus that can spread between people in healthcare settings [219]. This detailed clinical work accompanied with the identification of the virus clusters using NGS helped to identify the source of infection in the eastern region of Saudi Arabia. This discovery aided with NGS had immediate implications in terms of preventive infection control measures to halt the spread of the virus to other regions of the world.

Non-invasive diagnosis and monitoring

Detecting foreign DNA from the blood is an example of a novel NGS application. NGS platforms have been used to monitor solid-organ transplant rejection by detecting cell-free DNA from the blood [220]. The ratio of recipient genomic DNA to graft-derived donor DNA is used to measure the number of graft cells that are dying and releasing their DNA into the blood. This method has a big advantage of being less invasive compared with traditional methods requiring periodic biopsies of the graft tissue.

Similarly, prenatal diagnosis of several trisomies is now possible with NGS without the need for traditional invasive amniocentesis. Here, NGS are used to sequence cell-free DNA from the maternal blood in order to detect fetal trisomies by comparing the ratios of the number of DNA fragments derived from each chromosome [221]. This technique showed impressive records of sensitivity and specificity of detection of fetal trisomy 21, 100% and 97.9% respectively [222].

Population genetics

The 1000 genomes project (1KG) is probably one of the most notable NGS applications [155, 197]. The 1KG used both low-coverage whole genome sequencing and exome sequencing of 1,092 individuals from 14 populations to provide a haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. The 1KG captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% and provides a valuable resource in many projects including population frequency-based filters used in the exome sequencing projects analyzed in this thesis.

Another influential study entailed high-coverage exome sequencing of 6,515 individuals [199]. The study shows that 73% of all protein-coding SNVs and approximately 86% of SNVs that are predicted to be deleterious, arose in the past 5,000–10,000 years. Additionally, it identified an excess of rare coding mutations in essential and Mendelian disease genes in Europeans compared to African Americans, a finding consistent with weaker purifying selection due to the smaller effective population sizes resulting from the Out–of –Africa dispersal.

Forensics

DNA-based methods for human identification are generally based on genotyping of short tandem repeat (STR) loci using electrophoresis, which is relatively low throughput and does not yield nucleotide sequence information. NGS platforms have been used as high-throughput genotyping analysis for the 13 Combined DNA Index System (CODIS) STR loci and amelogenin (AMEL) locus using as few as 18,500 reads (>99% confidence) [223]. STRait Razor is a program developed to detect forensically relevant STR alleles in FASTQ sequence data, based on allelic length. Currently, it detects alleles for 44 autosomal and Y-chromosome STR from Illumina sequencing instruments with 100% concordance [224].

Functional applications

NGS has many applications that extend outside the scope of genome sequencing. The ENCODE project demonstrates the breadth of various non-genome-based NGS experiments (Table 1-11). In this project, a total of ~1659 high-throughput experiments were performed to analyze transcriptomes and identify methylation patterns in human genome [225]. This is a large multicenter project has assigned biochemical activities to 80% of the genome, particularly the annotation of non-coding portions in the genome [226]. This finding may help to improve the prioritization and interpretation of non-coding variants frequently found in whole genome sequencing project.

Table 1-11 The various NGS assays employed in the ENCODE project to annotate the human genome [226]. HT: high-throughput

Feature	Method	Description	Reference
Transcripts, small RNA and transcribed regions	RNA-seq	Isolate RNA followed by HT sequencing	[227]
	CAGE	HT sequencing of 5'-methylated RNA	[228]
	RNA-PET	CAGE combined with HT sequencing of poly-A tail	[229]
	ChIRP-Seq	Antibody-based pull down of DNA bound to lncRNAs followed by HT sequencing	[230]
	GRO-Seq	HT sequencing of bromouridinated RNA to identify transcriptionally engaged PolII and determine direction of transcription	[231]
	NET-seq	Deep sequencing of 3′ ends of nascent transcripts associated with RNA polymerase, to monitor transcription at nucleotide resolution	[232]
	Ribo-Seq	Quantification of ribosome-bound regions revealed uORFs and non-ATG codons	[233]
Transcriptional machinery and protein–DNA interactions	ChIP-seq	Antibody-based pull down of DNA bound to protein followed by HT sequencing	[234]
	DNAse footprinting	HT sequencing of regions protected from DNAse1 by presence of proteins on the DNA	[235]
	DNAse-seq	HT sequencing of hypersensitive non-methylated regions cut by DNAse1	[236]
	FAIRE	Open regions of chromatin that is sensitive to formaldehyde is isolated and sequenced	[237]
	Histone modification	ChIP-seq to identify various methylation marks	[238]
DNA methylation	RRBS	Bisulfite treatment creates C to U modification that is a marker for methylation	[239]
Chromosome-interacting sites	5C	HT sequencing of ligated chromosomal regions	[240]
	ChIA-PET	Chromatin-IP of formaldehyde cross-linked chromosomal regions, followed by HT sequencing	[241]1

1.2.3 NGS challenges

Recent advances in NGS technologies have brought a paradigm shift in how researchers investigate human disorders. The key advantage of NGS is their ability to generate vast amount of biological data in a short time frame and in a cost-effective way. Despite their huge success, they are not without challenges. These challenges include *in silico* analysis, data privacy, data interpretation and ethical considerations.

The amount of data that NGS platforms generate can be unmanageable in terms of data storage and processing. The cost of sequencing a base is dropping faster than the cost of storing a byte [242]. Another issue caused by this large amount of data is that statistical analysis and data processing (e.g. imputation) of few hundreds to thousands of exomes or genomes can be very computationally intensive and almost always requires a large infrastructure of distributed servers, which may not be affordable for many researchers.

There are growing concerns with data privacy and whether current measures of sample anonymization are sufficient. It has been reported that a minimum number of 75 independent SNPs, or fewer, will uniquely identify a person [243]. It is even possible to re-identify genotyped individuals or even individuals in pooled mixtures of DNA [244]. This prompted the National Institute of Health (NIH), the Broad Institute in the US, and the Wellcome Trust in the United Kingdom to further restrict public access to the data from genome-wide association studies [245].

The biological and clinical interpretation of genetic variation is probably one of the remarkable challenges in the era of NGS. Most of the variants found in whole-genome sequencing are non-coding and many of the coding ones are of variants of uncertain significance (VUSs) [246]. Functional studies are required to evaluate these VUSs properly but with tens or hundreds of coding VUSs per individual, this is clearly is not a scalable solution.

At the ethical level, NGS raises many important questions. For example, when to return results to participants, and what are the researcher's obligations, if any, towards the participants' relatives. Such ethical dilemmas are the subject of heated debate between researchers, clinicians and policy makers [247] and are being actively addressed.

1.3 Overview of the thesis

In this thesis I establish an analytical infrastructure for exome sequence analysis and apply it to some simple monogenic scenarios where linkage analysis is used to guide the targeted NGS sequencing. I then apply it to two subtypes of CHD exploring the power of different study designs.

Chapter 2 describes the development of an analytical infrastructure and the workflow used to analyze exome data in family-based study designs. First, I describe two pipelines used to call variants in all samples analyzed in this thesis in addition to a third pipeline that I designed and implemented to call *de novo* variants. Variants called by these pipeline were subjected to various quality control tests and additional filters to improve the sensitivity and specificity of the variant calling. I then explain how the number of candidate genes per exome varies in different family designs and also by utilizing different public resources of minor allele frequency (MAF). To automate many of these analytical steps, I developed a suite of tools called Family-based Exome Variants Analysis or (FEVA) to report candidate genes in different study designs. FEVA has two interfaces: one is aimed to users without bioinformatics training (with a graphical user interface) while the other is a command-line interface suitable for high-throughput settings in large-scale projects. Finally, I present several applications on how I used FEVA to identify candidate genes in different study designs that include linkage regions in index cases, affected sib-pairs, trios, and affected parent-child pairs. The tools and analytical strategies described in this chapter were used to explore the power of different study designs in two CHD subtypes in the subsequent chapters.

Chapter 3 describes how exome sequencing combined with tools developed in chapter 2 were used to report *de novo* and recessively inherited variants in 30 trios with Tetralogy of Fallot (ToF). This is followed by custom targeted sequencing of 122 genes in a replication cohort of 250 additional ToF trios. This chapter also describes three additional analyses that I designed and performed that are not described in chapter 2: a modified transmission disequilibrium test (TDT) to explore incomplete penetrance of rare coding variants, an analysis of digenic inheritance, and finally a pathway burden analysis.

Chapter 4 discusses an alternative study design where I combined the analysis from 13 trios and 112 index cases to discover a novel CHD gene in patients with Atrioventricular Septal Defects (AVSD). Beside *de novo* and recessively inherited coding variants, this chapter descries a new analysis not described in chapter 2 that aims to test for the burden of rare coding variants in case/control samples.

Concluding remarks and future directions are detailed in **Chapter 5**.

2 | Developing, testing and applying analysis pipelines for family-based exome studies

2.1 Introduction

Although a rare genetic disorder, by definition (according to the European Commission), has a frequency of 1 in 2000, collectively rare diseases affect 6-10% of the population [248]. Rare genetic disorders are associated with high mortality rates, may account for 51% of deaths in children under 1 year [249], add a significant burden to the health care system in terms of cost (accounted for 184% more hospital charges than children who were hospitalized for other reasons [250]) and often under diagnosed [251].

Studying rare genetic disorders is essential to improve the quality of health care services and to obtain a precise and early diagnosis to these patients. Additionally, the insights from rare genetic disorders have helped to improve our understanding of many novel genes and molecular phenomena such as uniparental disomy, parental imprinting and epistatic interactions. These insights have also improved our understanding of the etiology of the risk and pathology of complex disease. For example, studying severe forms of familial insulin resistance has revealed important key genes when studying the common form of Diabetes Mellitus Type II [252].

In the last few decades, researchers have used different approaches to find the underlying genetic causes of rare disorders, such as positional cloning, linkage analysis and candidate gene resequencing among other methods. Despite these great efforts, the Online Mendelian Inheritance in Man (OMIM) [253] database lists 3,675 suspected Mendelian phenotypes without any known molecular basis , as of January 7th 2013. This large number of unidentified disorders shows the limitation of the traditional tools in identifying their genetic causes.

Next Generation Sequencing (NGS) platforms promise to accelerate this process. In 2005, the 454 Roche sequencer was introduced to the scientific community and soon other similar platforms followed, such as the Genome Analyzer from Illumina, SOLiD from Life Technologies and many others (discussed in chapter 1). These NGS platforms are able to generate unprecedented high-throughput DNA sequencing from whole genome or targeted sequences (e.g. exome or linkage regions) in a very short time and at an affordable cost. The first successful example of finding causal variants in a novel gene was published in 2010 when Sarah Ng *et al.* [174] used NGS to sequence the whole exome of four patients with Miller syndrome (OMIM #263750) and showed that mutations in the *DHODH* gene cause this recessive disorder. Soon afterwards, other groups around the world started using NGS to discover the causes of more than 100 novel genes in less than 3 years (Figure 2-1). This number is expected to grow as more researchers adopt NGS platforms for gene discovery in other monogenic disorders [202, 254] (discussed in monogenic disorder section in chapter 1).

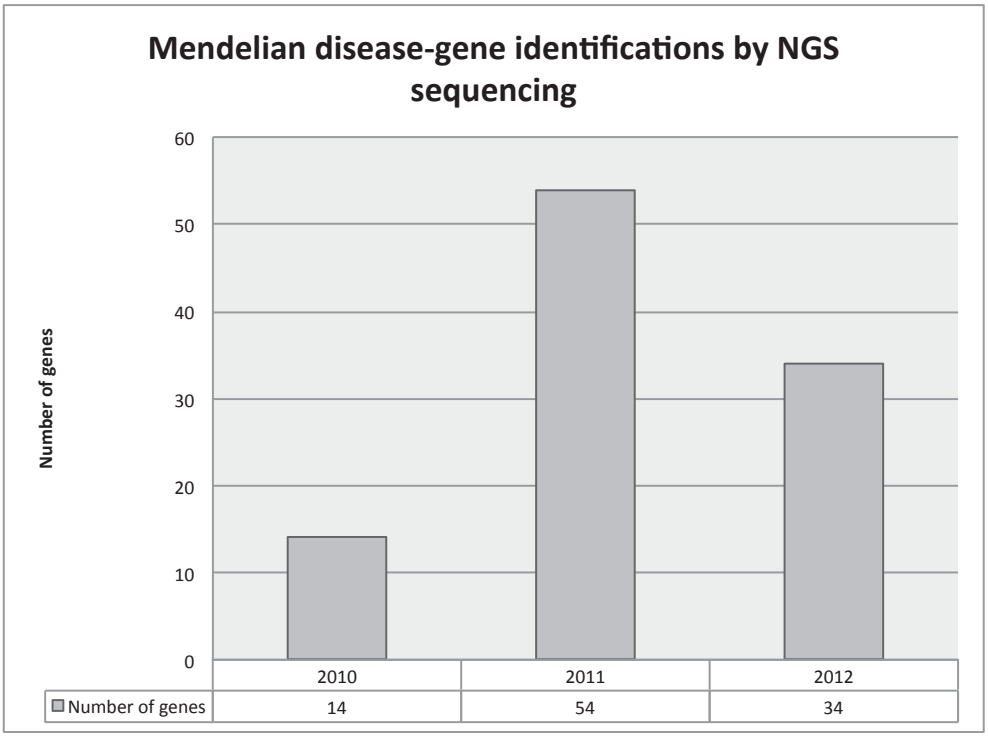


Figure 2-1 Number of Mendelian disease genes identified by NGS 2010 to mid of 2012 [254]

Congenital heart defects (CHD) are considered the most common birth defects worldwide when taken collectively [14]. However, they are considered rare disorders when considered separately (CHD prevalence is review in chapter 1). Inspired by the success of NGS in finding the genetic causes in other rare disorders, I approached CHD using family-based study designs combined with NGS.

However, since the genetic architecture of CHD is not currently clear, I have considered both Mendelian and non-Mendelian contributions to CHD. Not all pathogenic mechanisms can be evaluated using exome sequencing since it targets a small proportion of the genome (only coding DNA regions or < ~1-2% of the human genome size (Table 2-1). Cryptic splice sites, intragenic and long-range promoter variants that affect gene regulation cannot be studied using exome sequencing alone, and as such as they do not fall within the scope of this thesis. The existing examples of genetic causation of CHD are diverse, with respect to both their modes of inheritance and molecular mechanisms, and so investigation of CHD by exome sequencing requires a suite of tools capable of exploring different scenarios.

Table 2-1 lists the major inheritance patterns with syndromic or / and isolated CHD examples from literature, and whether they are amenable to analysis in whole exome sequence data (WES) or not, using tools I developed or implemented to scrutinize the candidate variants.

Table 2-1 Selected patterns of Mendelian and non-Mendelian inheritance and whether they are amenable to analysis using whole exome data. * Indicates mechanisms that have been evaluated in this thesis.

	Inheritance pattern	Example of syndromic and/or isolated CHD	Can be evaluated with WES?	Software	Explored in this thesis?
Mendelian	Autosomal recessive *	Adams-Oliver syndrome OMIM # 100300	Yes	FEVA	Chapter 2, 3 and 4
	Autosomal Recessive (compound heterozygous) *	five affected children with right atrial isomerism were compound heterozygotes for truncating mutations in *GDF1* gene [255]	Yes	FEVA	
	Autosomal dominant *	Alagille syndrome OMIM # 118450	Yes	FEVA	
	X-linked dominant *	Opitz GBBB syndrome OMIM # 300000	Yes	FEVA	
	X-linked recessive *	X-linked heterotaxy OMIM # 306955	Yes	FEVA	
	Y-linked	No reported CHD cases. Unlikely to harbor heart developmental genes	Yes	FEVA	Not explored
Non-Mendelian*	Recurrent *de novo* mutations *	*De novo* mutations in histone-modifying genes in isolated and syndromic CHD cases using exome data [256]	Yes, if in coding regions	DenovoGear	Chapter 3 and 4
	Digenic inheritance *	No reported CHD cases. But as an example: long QT syndrome	Yes	Digenic module	Chapter 3
	Polygenic inheritance	Tetralogy of Fallot [257]	Only with large sample size (in thousands), case/control analysis	Case/Control analysis	Not explored
	Imprinting	Prader-Willi syndrome OMIM # 176270 [258]	Yes, if large segment.	Uniparental Disomy (UPD) caller by Dan King,	Not explored
	Excess affected cases (segregation distortion) *	*MTHFR* C677T polymorphisms may contribute to the risk of CHDs [259]	Yes, in trio based studies	Rare collapsed TDT module	Chapter 3

2.1.1 Chapter overview

The main goal of this chapter is to describe the pipelines and analytical tools I developed and then applied to evaluate the utility of four family-based study designs (index cases with linkage analysis, affected sib-pairs, trios and affected parent-child). The lessons learnt from these analyses were subsequently applied to two CHD subtypes (Tetralogy of Fallot and Atrioventricular Septal Defects) in chapters 3 and 4, respectively. Figure 2-1 shows the main analytical components required for family-based exome studies.

In this chapter, first, I describe the **three pipelines used to call SNVs and indels** from all CHD samples included in this thesis. My colleagues at the Wellcome Trust Sanger Institute implemented two of the three pipelines (the Genome Analysis Production Informatics (GAPI) and the (UK10K) pipelines whilst I implemented the third one to call *de novo* variants, which was later adapted by Ray Miller for the Deciphering Developmental Disorders (DDD) project [260].

Each pipeline outputs a large number of variants including many false positive variants that would adversely affect any downstream analysis. At the beginning of my work on exome sequencing three years ago, it was not clear what best practices I should use to **improve the sensitivity and specificity** of variant calling. In the second part of the results, I describe how I chose **various filters** such as strand bias, phred-like quality scores among other filters to improve the sensitivity and specificity of the variant calls. Choosing the right filters is a dynamic research area and the best practices are expected to change to reflect new statistical models for variant calling. Many of the results I describe in this section do not reflect the current best practices but they represent examples of how to approach and set proper filter thresholds in exome-based studies. In addition to these filters, I discuss how I merged the variant calls from **multiple callers** to enhance sensitivity. I show that the precise manner in which the outputs from these callers are combined can have an unexpectedly large effect on the number of candidate variants

Once I have obtained a high quality set of variants for each sample, I describe in the third part of the results, how I used minor allele frequency and additional family data to minimize the search space for causal variants. These combined steps reduce the search space for causal variants to a few tens or hundreds instead of tens of thousands of variants.

Finally, I describe a suite of tools that I have designed to automate many steps discussed above. Although similar software, such as SVA, EVA and VarSift [261-263], have been published during my PhD, none of them were able to fulfill the

needs for my studies. One of the main drawbacks of these tools is that they are not suitable for high-throughput analysis. Additionally, most of them use hard coded filters, which is not practical to explore new filters. For these reasons, I developed a suite of tools called **Family-based Exome Variants Analysis (FEVA)** that reports candidate variants under different modes of inheritance (autosomal recessive, autosomal dominant and X-linked) for different study designs (index cases, affected sib-pairs, affected parent-child, and trios). In the last part of this chapter, I show how I used FEVA to identify pathogenic and candidate pathogenic genes under different study designs using real examples.

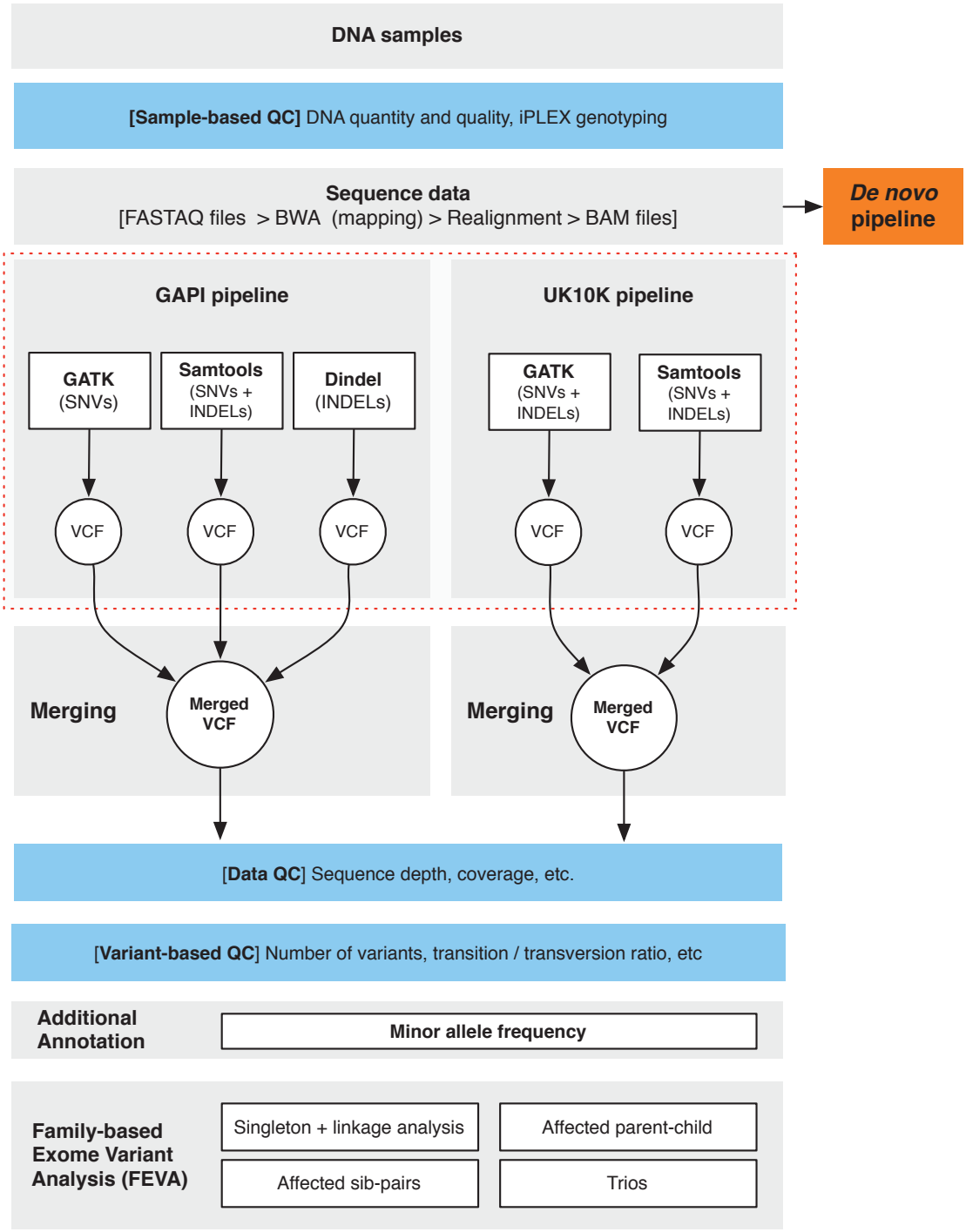


Figure 2-2 Overview of pipelines, tools and annotation discussed in this chapter.
Blue boxed are quality control tests that are performed at different stages of the workflow. The two main pipelines used to call variants from sequence data are GAPI and UK10K. A third one, the de novo pipeline (orange box), uses the sequence data (BAM files) and includes further steps described in Figure 2-9. Additional descriptions of these steps are available in Table 2-2.GAPI: the Genome Analysis Production Informatics pipeline, UK10K: UK10K variant calling, SNVs: single nucleotide variants, INDELs: insertion and deletion, QC: quality control.

Table 2-2 A list of main analytical tasks described in this chapter with a short description of each section.

Task	Section	Description
Variant calling pipelines	Genome Analysis Production Informatics (GAPI) pipeline	To call single nucleotide (SNVs) and insertion/deletion variants (INDELs) using three callers (Samtools, GATK and Dindel) in 381 CHD samples
	UK10K pipeline	Used to call SNVs and INDELs variants using two callers (Samtools and GATK) in 125 CHD samples.
	De novo variant calling pipeline	Used to call *de novo* SNVs and INDELs variants using one caller (DenovoGear) in 252 CHD trios
Improving sensitivity and specificity	Sample-based DNA quality test (DNA samples)	Various tests to detect the quantity and quality of the DNA samples and any possible sample contamination and swapping issues.
	Sample-based data quality test (Sequencing data)	Quality of NGS sequencing data in terms of depth, coverage and other parameters.
	Variant-based quality tests	Quality of variant calling based on the number of variants, genotypes, variants predicted effect on the protein and other quality ratios.
	Filtering low quality variants	Multiple filters based on thresholds of quality metrics used to exclude low quality variants
	Using multiple callers	Combining multiple variant callers (e.g. Samtools, GATK and Dindel) to overcome the deficiencies of individual callers
Minimizing the search space for causal variants	Minor allele frequency (MAF)	Using different population-based MAF resources to exclude common variants (>1%) and the effect of allele matching algorithm.
	Family-based designs	The effect of considering additional members of the family (either healthy or affected) on the final number of candidate variants and genes
Applications	FEVA suite	An easy to use suite of programs I developed to automate many of the steps discussed above (minimize the search space for causal variants and prioritization). These tools are available for small scale use with a graphical user interface and as common-line tools for high-throughput analysis.
	Simple monogenic diseases combined with linkage analysis	Use of FEVA to find pathogenic variants from four different index cases within linkage intervals for different neurodevelopmental monogenic disorders
	Affected sib-pairs	Use of FEVA to analyze CHD in affected sib-pairs from eight non-consanguineous and two consanguineous families.
	Affected parent-child	Using FEVA to analyze CHD in three affected parent-child pairs.
	Example of affected trios combined with candidate gene screening	Use of FEVA to analyze 1,080 trios from Deciphering Developmental Disorders (DDD) project trios and screen 1,142 candidate genes.

2.2 Methods

2.2.1 Samples and phenotypes

Table 2-3 summarises the different sample collections that I analyzed to evaluate the utility of different study designs. These sample collections were accessed through collaboration with various researchers and clinicians from the UK, Europe and Canada. All samples were collected from the families after obtaining informed consents and approved by the Ethical Review Boards of their respective organizations. Not all of the analyses of these sample sets are described in detail in this thesis.

Table 2-3 Samples and family-based study designs included in this thesis.
* Sample cohorts discussed in this chapter. GO-CHD: Genetic Origins of Congenital Heart Disease Study, DDD: Deciphering Developmental Disorders project, AVSD: atrioventricular septal defects. TOF: tetralogy of Fallot.

Design	Targeted Region	Cohort	Origin	Consanguineous	Phenotype	Number of families or samples
Index cases	Whole exome	GO-CHD	UK	No	Various CHD	110
		Toronto	Canada	No	AVSD	78
	Linkage region	Amish*	USA	No	Various Neurodevelopmental	4
Trios	Whole exome	GO-CHD	UK	No	Various CHD	2
		Newcastle	UK	No	TOF	30
		Toronto	Canada	No	AVSD	3
		Leuven	Belgium	No	AVSD	10
		DDD	UK	No	Developmental	1,080
	Candidate genes	Newcastle	UK	No	TOF	250
Affected sib-pairs	Whole exome	Toronto	Canada	No	AVSD	1
		Birmingham*	UK	Yes	Various CHD	2
		Birmingham*	UK	No	Various CHD	8
		GO-CHD*	UK	No	Various CHD	1
Affected parent-child	Whole exome	GO-CHD*	UK	No	Various CHD	3

2.2.2 DNA preparation and Quality Control

Our collaborators extracted the DNA from the patients' blood and / or saliva and sent the samples to the Sanger Institute for quality control before they were submitted for sequencing. The DNA sample quality control included three tests. The first was to determine the amount and concentration of DNA, which was analyzed by gel or picogram. The second test detected the sample's gender by genotyping SNPs on the sex chromosomes and compared it to the supplier sheet in order to detect any potential gender mismatches. The third test was to check for the possibility of sample contamination or swapping by genotyping another 30 SNPs. The genotyping was done using Sequenom platform and any sample, which failed one of these tests, was flagged for replacement or exclusion. The Sample Logistic Team at the Sanger Institute performed these quality control tests.

2.2.3 Target capturing and sequencing

DNA (1-3µg) was sheared to 100-400 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation and enriched for target sequences (Agilent Technologies; Human All Exon 50 Mb - ELID S02972011) according to manufacturer's recommendations (Agilent Technologies; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). Enriched libraries were sequenced using the HiSeq platform (Illumina) as paired-end 75 base reads according to manufacturer's protocol.

2.3 Results

2.3.1 Assessing variant calling pipelines

2.3.1.1 Genome Analysis Production Informatics (GAPI) and UK10K pipelines

There are several pipelines deployed at the Wellcome Trust Sanger Institute (WTSI) to call variants from human whole genome and / or whole exome data. The majority of samples analyzed in this thesis were processed through the Genome Analysis Production Informatics (GAPI) pipeline (managed by Carol Scott *et al.*) except 125 samples that formed part of the UK10K RARE project, which were processed through the UK10K pipeline (managed by Shane McCarthy *et al.*) [264]. Both pipelines are used to call single nucleotide variants (SNVs) as well as insertion/deletion variants (INDELs). The GAPI pipeline provided single-sample calling only while UK10K pipeline provided both single and multi-sample calling. Although, the latter has some potential advantages, I decided to use single-sample calling only in order to be able to compare variants from both pipelines.

However, differences between these pipelines led to variability in the type and numbers of variants (Table 2-4, Table 2-5 and Figure 2-5). Data that were processed through the GAPI pipeline tend to have a larger number of SNVs and INDELs compared to UK10K pipeline. GAPI sequence data had 60% more SNVs compared with UK10K data although most of these differences can be attributed to non-coding variants which include intronic, intragenic, downstream, upstream and variants in untranslated regions UTRs).

To see if using different filters and thresholds in Table 2-5 caused the difference seen in SNVs counts between the two pipelines, I applied UK10K's filters on samples from the GAPI pipeline. First, I created a new set of samples called GAPI-II by merging variants from GATK and Samtools only and excluding Dindel calls since it is not part of the UK10K pipeline. This set of samples showed a similar number of coding and non-coding variants between both pipelines (Figure 2-4)

except for loss-of-function variants (LOF) where the UK10K pipeline has almost double the number of LOF variants compared with GAPI or GAPI-II (t test, *P* value < 2.2×10^{-16}). A difference in a caller version and its underlying statistical model is likely to cause this variation. This is more readily observed in LOF counts since they are fewer than missense variants and have a lower number of true variants and so are more sensitive to calling errors.

On the other hand, INDELs show larger differences between GAPI and UK10K pipelines (Figure 2-5). GAPI calls almost two to three times more INDELs than UK10K or GAPI-II (Figure 2-5-A). This is true regardless of the location of the indel with respect to coding sequences (Figure 2-5 B, C and D). One explanation for this observation would be the use of an additional caller specifically designed to call INDELs, called as Dindel, in the GAPI pipeline but not in the UK10K pipeline. Dindel is a dedicated caller for INDELs that uses a probabilistic realignment model to account for base-calling errors, mapping errors, and for increased sequencing error INDEL rates in long homopolymer runs [158]. Dindel's superior performance comes at a price of high computation demands, and the same underlying model has been incorporated into later versions of SAMtools, which is why the UK10K informatics team has refrained from using it on large numbers of samples.

Table 2-4 Similarities and differences between the components of Genome Analysis Production Informatics (GAPI) pipeline and the UK10K pipeline. Multiple factors are likely contribute to the differences in the number of variants generated by GAPI compared with UK10K pipeline such as the number of used callers, different software versions which usually reflect subtle changes in the underlying statistical models, filters and thresholds and how the output from different callers is merged (i.e. the order of callers from the most to least preferred, see section 2.3.2.2 for details)

Step	Goal / Description	GAPI	UK10K
Reference genome	Which version of the human reference genome used	GRCh37 (hs37d3) 1000 genome phase II reference	GRh37 (human_g1k_b37) 1000 Genomes Phase 1 reference
Align sequence reads to reference genome	Generate SAM/BAM files	BWA (v0.5.9-r16)	BWA (v0.5.9-r16)
Mark duplicates	To mitigate the effects of PCR amplification bias introduced during library construction.	Picard tools (v1.46)	Picard tools (v1.46)
Realignment around indels	Enhance variant calling	GATK (v1.4-15)	GATK (v1.1-5-g6f432841)
Base quality score recalibration	Recalibrate base quality scores of reads according to the base features (e.g., reported quality score, the position within the read)	GATK (v1.4-15)	GATK (v1.1-5-g6f43284)
Calling target region	Calling variants is limited to the coding regions plus variable flanking region	Exon bait regions plus or minus a 100bp window	Exon bait regions plus or minus a 100bp window
SNV calling	Single nucleotide variants calling programs	Samtools (v0.1.16) GATK (v1.0.15777)	Samtools (v0.1.17) GATK (v1.3-21)
INDEL calling	Insertion and deletion variants calling programs	Samtools (v0.1.16) Dindel (v1.01)	Samtools (v0.1.17) GATK (v1.3-21)
Variant predicted effect	The effect of variant on the protein is predicted by VEP	VEP 2.2 to 2.4	VEP 2.6 to 2.8
Caller merging	The order of which variants called by different callers are merged	Dindel > GATK > Samtools	GATK > Samtools
General filters	Filters applied during variant calling	See Table 2-5 for details	

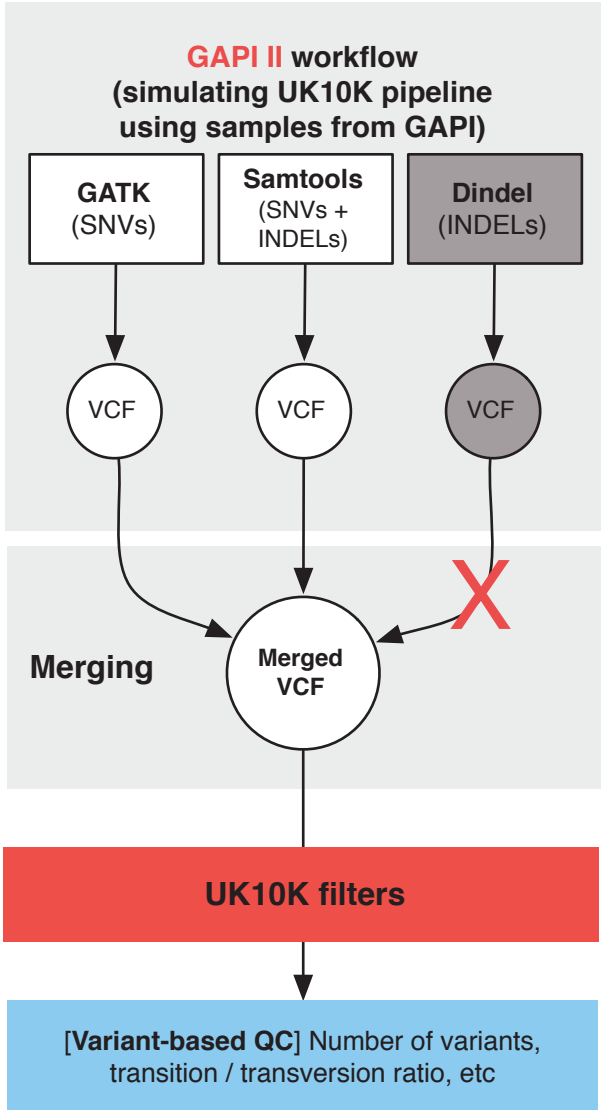


Figure 2-3 A workflow diagram to describe how I generated VCF files for GAPI-II set. The main goal is to use files from GAPI pipeline and apply similar workflow to UK10K and see if this would be enough to explain the differences between the pipelines.
Each sample from the original GAPI pipeline has three VCF files of variants called by GATK, Samtools and Dindel. I merged VCF files from GATK and Samtools but not from Dindel. Next, I applied the same filters used by UK10K to exclude low quality variants (filters were supplied by Shane McCarthy). A list of UK10K filters is available in Table 2-5.

Table 2-5 Filters and thresholds applied on variants from UK10K and GAPI pipelines.

Variant callers	Filters	Variant type	Pipelines	
			GAPI	UK10K
Samtools	Depth at locus (DP)	SNVs	4 < DP and DP > 1200	4 < DP and DP > 2000
		INDELs	4 < DP and DP > 1200	4 < DP and DP > 2000
	Mapping quality (MQ)	SNVs	MQ <=10	MQ <= 25
		INDELs	MQ <= 10	MQ <= 25
	Genotype quality (GQ)	SNVs	NA	GQ <= 25
		INDELs	NA	GQ <= 60
	Variant quality (QUAL)	SNVs	NA	QUAL <= 30
		INDELs	NA	QUAL <= 60
	StrandBiasPval	SNVs	StrandBiasPval < 0.0001	NA
		INDELs	StrandBiasPval < 0.0001	NA
	BaseqBiasPval	SNVs	BaseqBiasPval < 1e-100	NA
		INDELs	BaseqBiasPval < 1e-100	NA
	MapqBiasPval	SNVs	MapqBiasPval < 0	NA
		INDELs	MapqBiasPval < 0	NA
	EndDistBiasPval	SNVs	EndDistBiasPval < 0.0001	NA
		INDELs	EndDistBiasPval < 0.0001	NA
	MinbpfromGap	SNVs	MinbpfromGap < 10	NA
		INDELs	MinbpfromGap < 10	NA
GATK	Variant quality (QUAL)	SNVs	QUAL < 30	QUAL < 30
		INDELs	NA	NA
	Quality by Depth (QD)	SNVs	QD < 5.0	QD < 5
		INDELs	NA	QD < 2
	Homopolymer run length (Hrun)	SNVs	HRun > 5	Hrun > 5
		INDELs	NA	NA
	Strand bias (SB)	SNVs	SB > 10	SB > -0.1
		INDELs	NA	NA
	Fishers p-value (FS)	SNVs	NA	FS > 60
		INDELs	NA	FS > 200
	ReadPosRankSum	SNVs	NA	NA
		INDELs	NA	< -20
	InbreedingCoeff	SNVs	NA	NA
		INDELs	NA	< -0.8
	InDel	SNVs	Filtered if site covered by known indel mask file	Filtered if site covered by known indel mask file
		INDELs	NA	NA
	LowQual	SNVs	Repeat of QUAL < 30 (applied at calling)	NA
		INDELs	NA	NA
	SnpCluster	SNVs	Filtered if 3 SNPs within a 10bp window	NA
		INDELs	NA	NA
	Depth at locus (DP)	SNVs	4 < DP and DP > 1200	NA
		INDELs	4 < DP and DP > 1200	NA
	Hard to validate	SNVs	MQ0 >= 4 and (MQ0/(1.0*DP))	MQ0 >= 4 and (MQ0/(1.0*DP))
		INDELs	NA	NA
Dindel	Homopolymer run length (hp10)	INDELs	HRun > 10	NA
	Variant quality (q20)	INDELs	QUAL < 20	NA
	Non-reference allele (fr0)	INDELs	Not covered by at least one read on both strands	NA
	Multiple indels in the same window (wv)	INDELs	Other indel in window had higher likelihood	NA

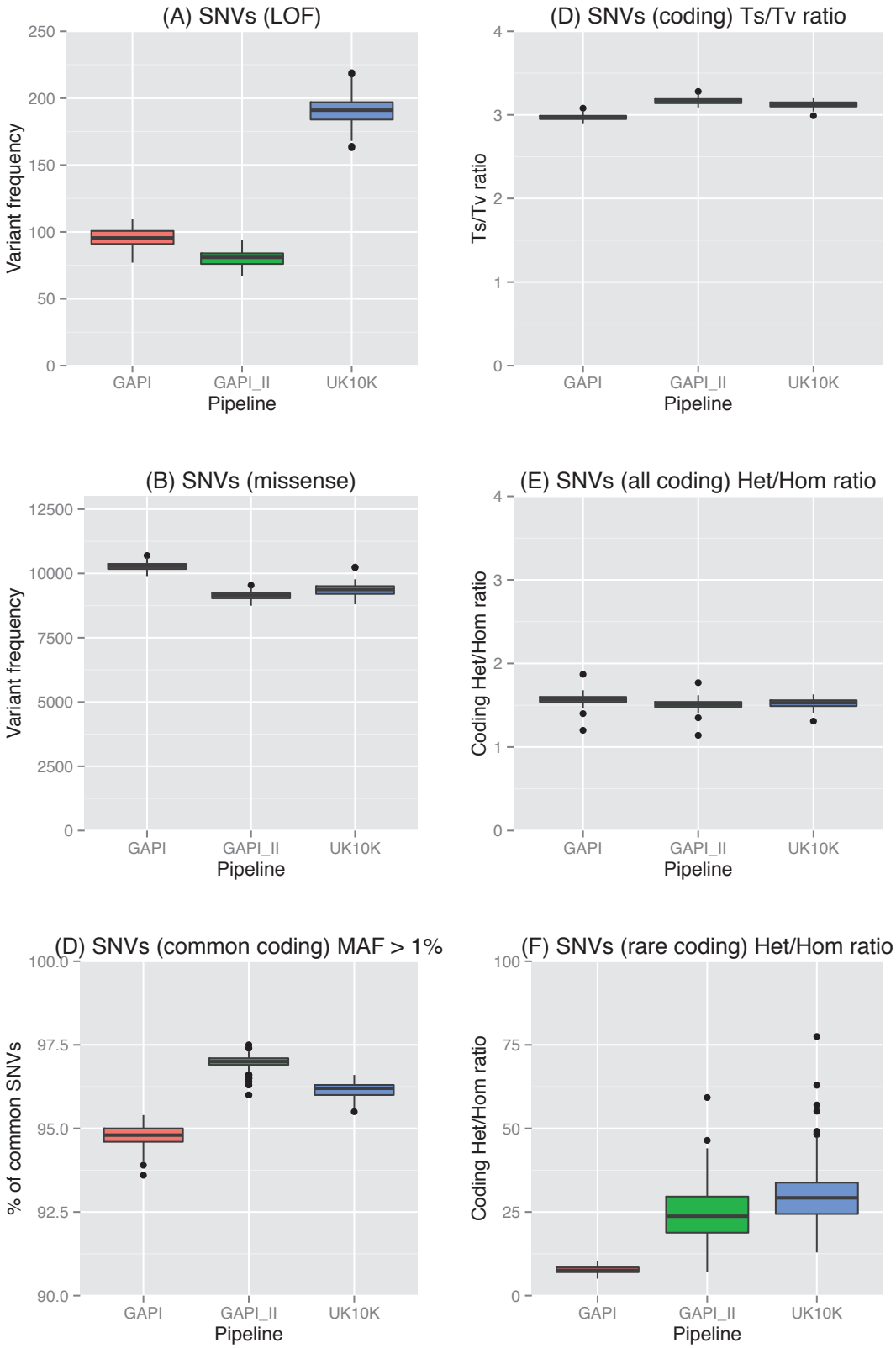


Figure 2-4 Differences in the counts of coding single nucleotide variant (SNVs) between GAPI and UK10K pipeline and GAPI_II, which include the same sample in GAPI but subjected to UK10K's filters (i.e. I applied the UK10K filter in Table 2-5 on GAPI samples).
LOF: loss-of-function variants include stop gain and variant disturbing donor or acceptor splice sites. Ts/Tv: Transition/Transversion ratio. Hom/Het: Homozygous/ Heterozygous ratio.

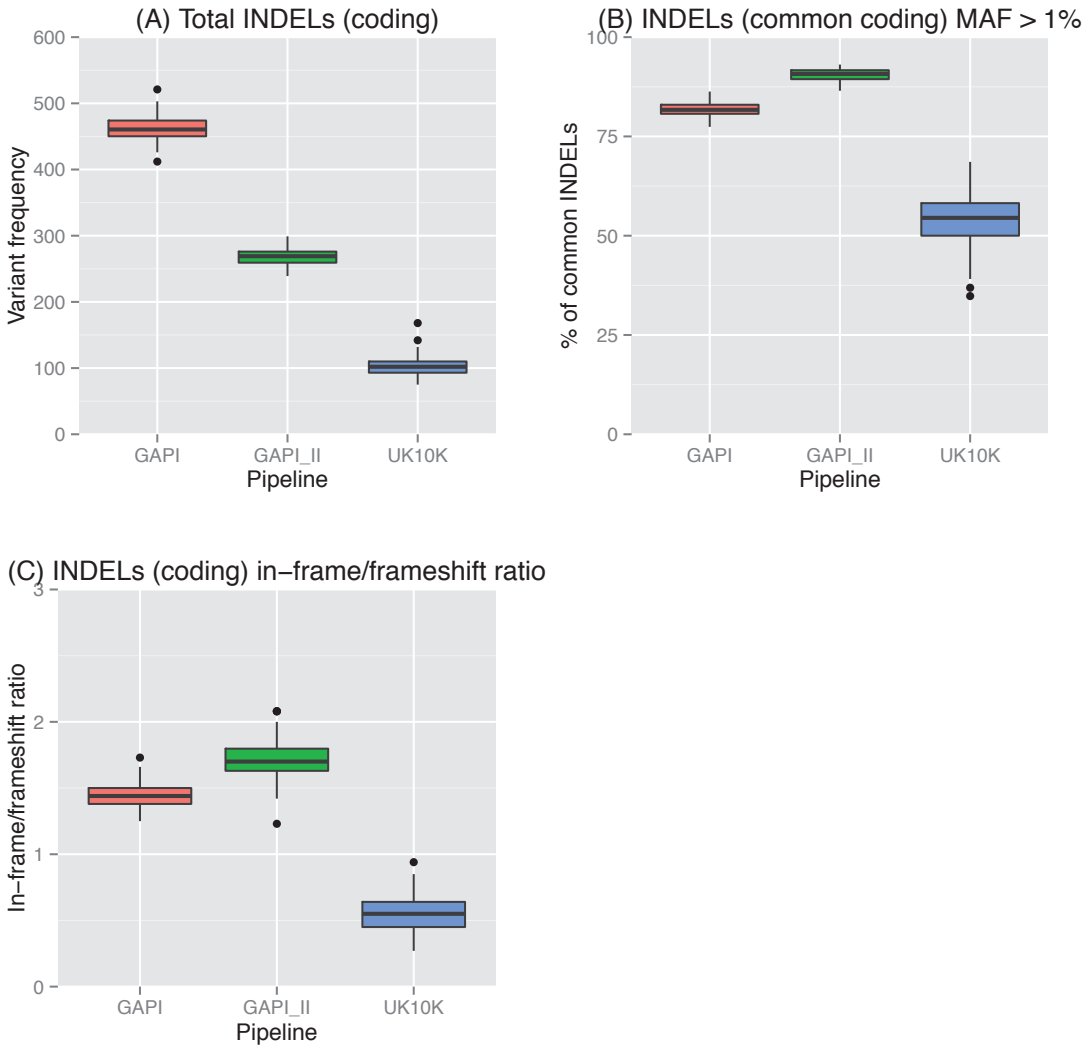


Figure 2-5 Differences of insertion-deletion variant (INDELs) counts between GAPI, UK10K pipeline and GAPI_II which are the same sample in GAPI but subjected to UK10K's filters).

2.3.1.2 Differences between GAPI releases

Since most of the samples analyzed in this thesis went through the GAPI pipeline at different points of my PhD, I sought to examine the effect of different releases of GAPI pipelines on the samples from three CHD cohorts (Figure 2-6 and Figure 2-7). The first cohort includes 94 samples of mostly atrioventricular septal defects (AVSD) children collected from SickKids hospital, Toronto, Canada (labeled as CHDT). The second cohort includes 90 samples of Tetralogy of Fallot (TOF) affected trios from the University of Newcastle while the third cohort includes 24 samples of affected sib-pairs of samples affected with various CHD

subtypes (about a quarter of these samples are from consanguineous families of a Pakistani origin). I found the variant counts were consistent between these cohorts even though they were generated at different times and with different versions of the GAPI pipeline. Small variations may occur as a result of systemic differences caused by the depth of the sequencing, or the population ancestry of the samples (e.g. samples with African ancestry are expected to have more variants than non-African samples).

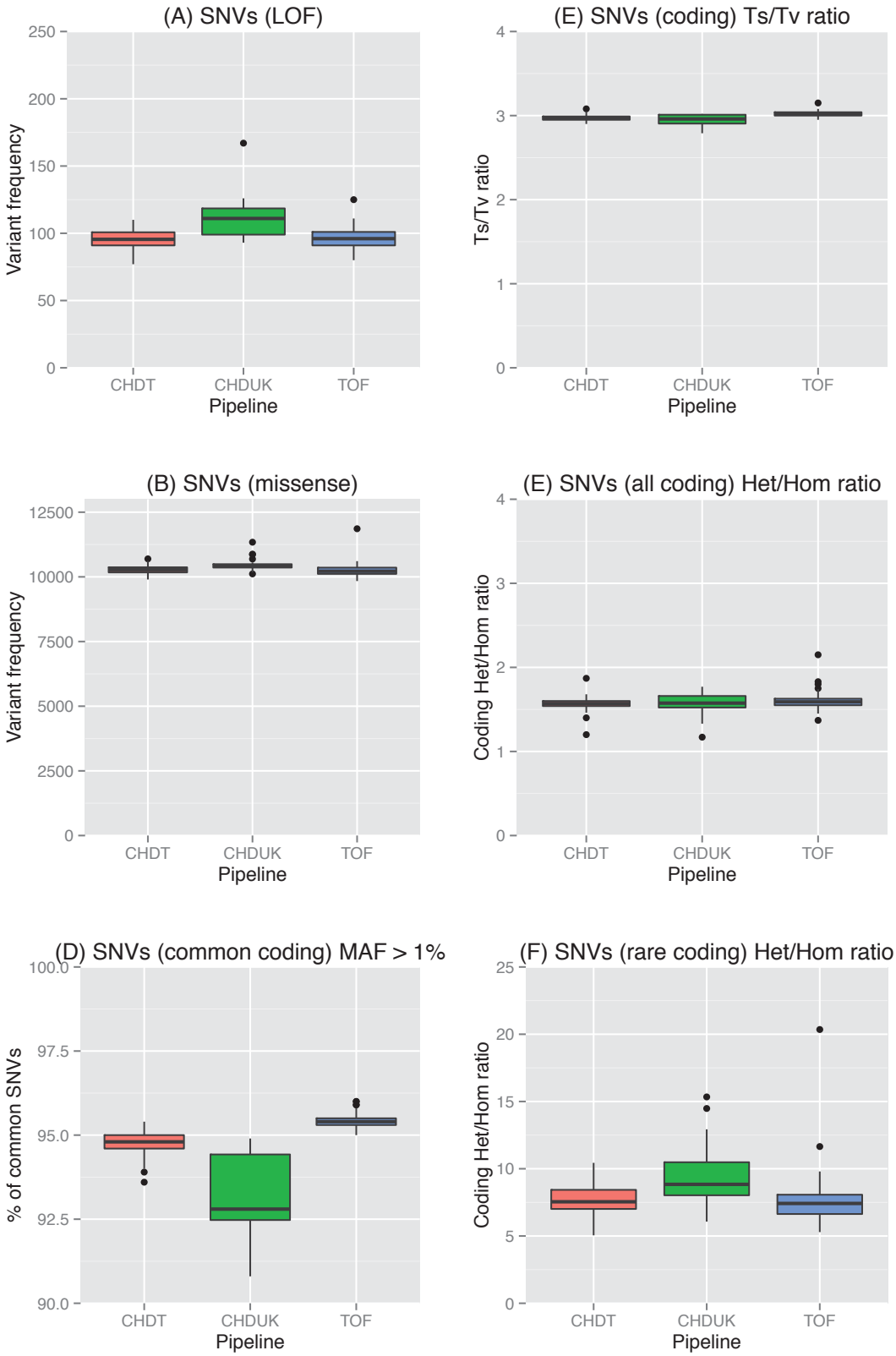


Figure 2-6 Differences of single nucleotide variant (SNVs) counts between GAPI studies. CHDT: Congenital heart defect samples from Toronto (discussed in chapter 4). CHDUK: Congenital heart defect samples from UK (discussed in application section in this chapter), TOF (Tetralogy of Fallot samples discussed in chapter 3). Ts/Tv: Transition/ Transversion ratio. Hom/Het: Homozygous/Heterozygous ratio.

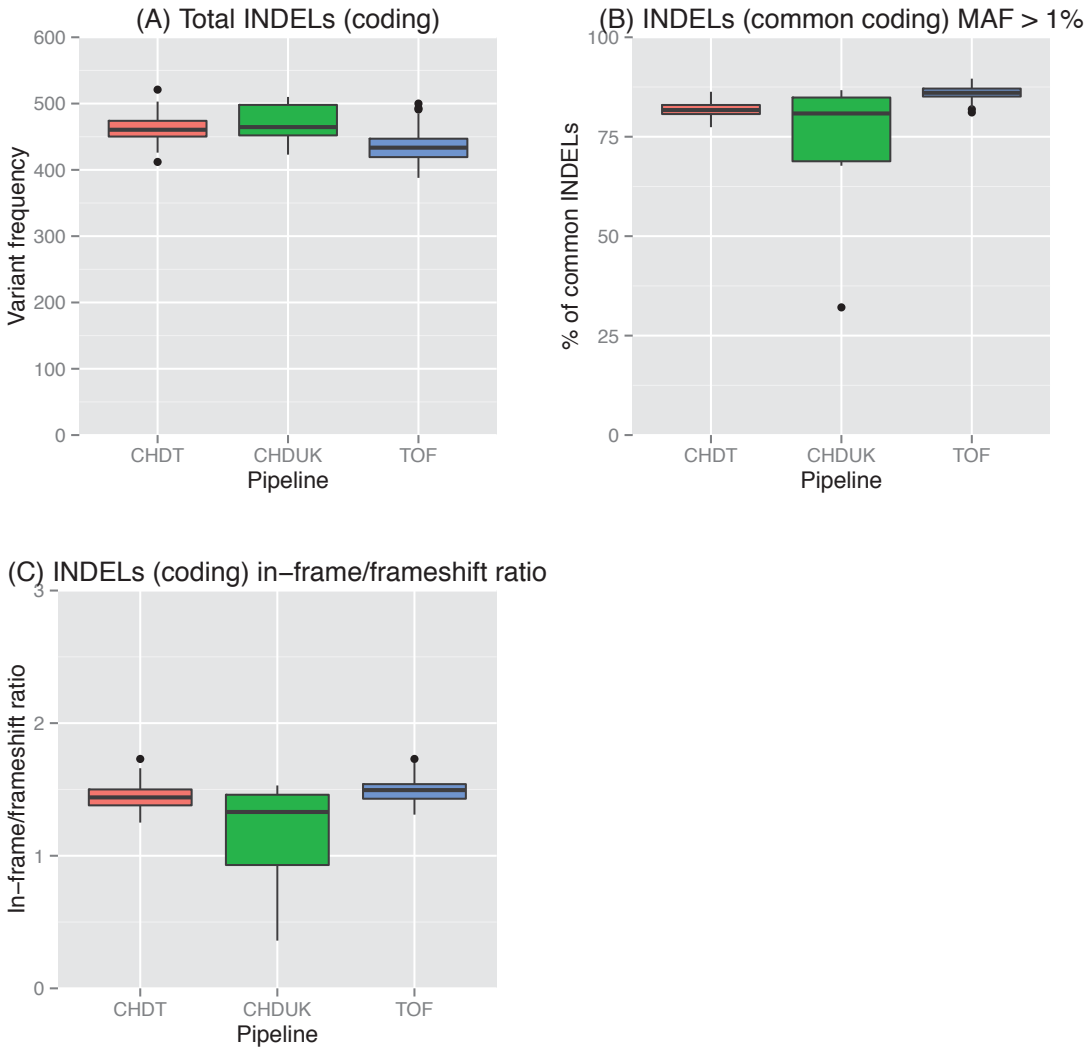


Figure 2-7 Differences of insertion-deletion variant (INDELs) counts between GAPI studies. CHDT: Congenital heart defect samples from Toronto (discussed in chapter 4). CHDUK: Congenital heart defect samples from UK (discussed in application section in this chapter), TOF (Tetralogy of Fallot samples discussed in chapter 3).

2.3.1.3 Implementing a de novo variant calling pipeline

Initially, I tried to identify potential *de novo* variants based on the variants called by either GAPI or UK10K pipelines in the child and not in parents. However, this approach yields a large number of candidate *de novo* variants per trio. A more efficient approach is to discover potential *de novo* variants from the child and his parents in a unified statistical framework. I designed and implemented a pipeline to call, filter, annotate and visualize *de novo* variants from trio-based studies based on DenovoGear program [265, 266]. This software was developed by Don Conrad and adopts a Bayesian approach to calculate the posterior probability of a *de novo* mutation at a single locus using the joint likelihood of the read-level data for all three trio members. DenovoGear outputs ~170 plausible *de novo* variants (with a posterior probability of greater than 0.001) per trio on average. However, most of these candidate variants are false positive since the expected number of *de novo* coding variants is ~1 according to published studies [190, 267-271].

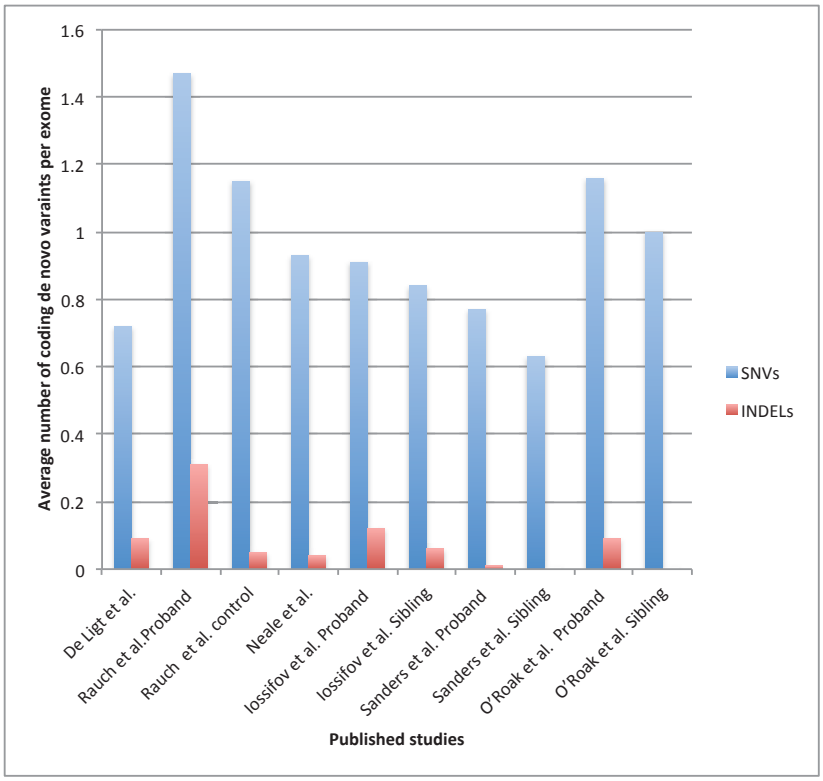


Figure 2-8 Average number of coding *de novo* variants per exome in different trio-based studies [190, 267-271]. (The literature survey and data are courtesy of Dr. Matthew Hurles)

In order to keep the number of false positive variants as small as possible, I applied five filters to exclude: (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 Twins cohort [264], (iii) when > 10% of the reads in either parent support the alternate 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 by the VEP tool [170]. Collectively, these filters effectively remove ~98.8% of the original candidate *de novo* variants (leaving ~1.8 coding plausible *de novo* candidate per exome).

This pipeline was used to automate several tasks designed to obtain high quality sets of candidate *de novo* variants from trios. This first step is calling candidate *de novo* variants from whole genome or whole exome data from human or mouse trio samples, followed by applying various filters to improve the specificity of the calls. The pipeline was designed in a modular fashion where each step generates intermediate files that are used as input for subsequent steps (steps are listed in Figure 2-9). This design allows the end user to change the pipeline by modifying steps and files or add new steps in order to customize the pipeline to suit the need of different studies.

One of the challenges faced by this pipeline is the run time per trio (~12 hours for whole exome data and up to 36 hours for whole genome data). To make the pipeline run faster, especially for large-scale project, I modified the code (which I wrote in Python programming language) to split sequence data in each sample into 24 segments (by the chromosome) and run them in parallel. This has shortened the run time to 2-3 hours for whole exome data and 10-12 hours for whole genome data. Moreover, another layer of parallelism is achievable by running multiple trios at the same time, which is suitable for large-scale projects such as the Deciphering Developmental Disorders (DDD) project with thousands of trios.

I used this pipeline to call *de novo* variants in 238 trios affected with Tetralogy of Fallot in the third chapter and in 13 trios with atrioventricular septal defect in the fourth chapter. Moreover, this pipeline has been used successfully in several whole genome sequencing projects in human and mouse pedigrees that are investigating the factors influencing rates of germline mutation.

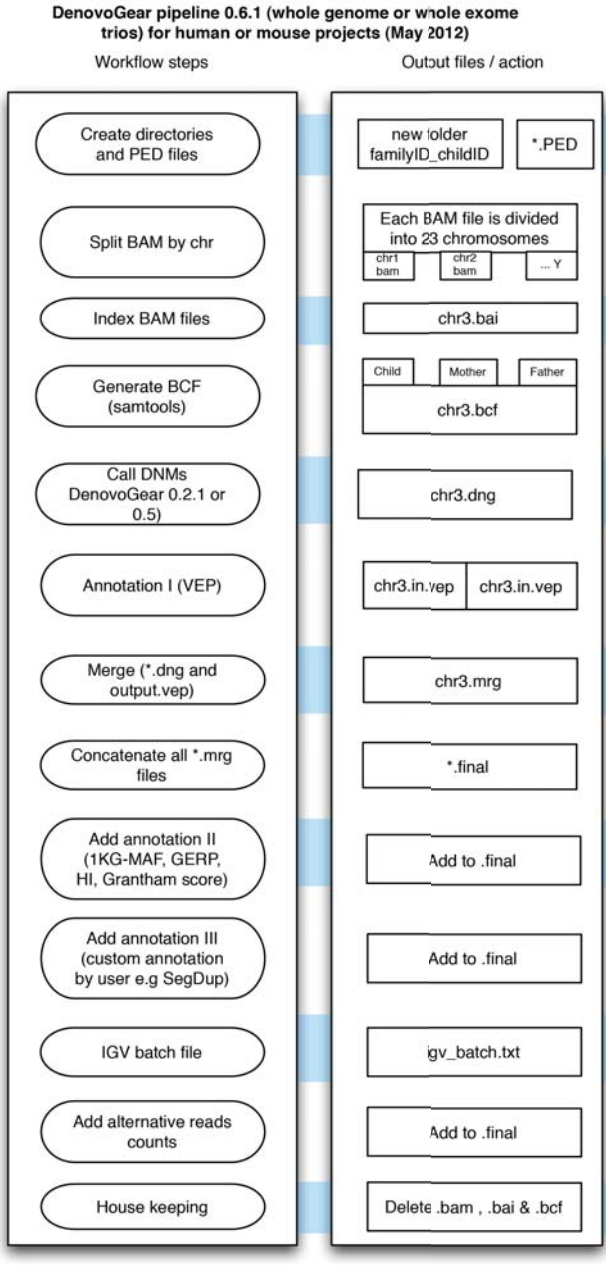


Figure 2-9 **The workflow of the DenovoGear pipeline.**
PED: pedigree files. BCF are binary files of VCF (variant call format) that are generated by Samtools mpileup with genotype likelihoods required by DenovoGear [272]. DNMs: *de novo* mutations. VEP: variant effect predictor [170]. 1KG-MAF: 1000 genomes minor allele frequency. GERP: Genomic Evolutionary Rate Profiling scores [164]. HI: haploinsufficiency scores [273].

2.3.2 Minimizing the rate of false positive variants

2.3.2.1 Variant-based filters

At the beginning of my PhD studies, it was not clear what were the best practices I should use to improve the sensitivity and specificity of variant calling from exome data. To investigate this aspect of data analysis, I tested different filters in order to determine the best callset possible from CHD samples called by the UK10K pipeline. These callsets include raw unfiltered variants called by GATK (G), Samtools (S), or both callers (GS). In this analysis, I focused mainly on SNVs since they are the most abundant variants and represent a large proportion of the known pathogenic variants [274]. More importantly, there are many high quality training SNVs data sets available to improve variant quality (e.g. HapMap). On the other hand, indels were, and still are, more difficult to call and tend to have a higher false positive rate [155].

SNVs are thought to be among the easiest variant classes to call from NGS data but nonetheless sequencing errors can generate false positive calls. Sequencing error rates depend on factors such as the context of the DNA sequence, depth of sequencing, and the type of substituted bases among other factors [143]. To control for these biases in the exome NGS data, I examined the relationship between strand bias (SB), quality by depth (QD), genotype quality (GQ) and variant quality (QUAL) with transition/transversion ratio (Ts/Tv). This ratio has been used by different groups in the 1000 genomes consortium as a quality control test and typically ranged between 2.9-3.3 in coding regions based on sequence data from different NGS platforms. I used the Ts/Tv ratio as the truth measurement to determine the proper thresholds values for each one of the four filters.

Variant quality (QUAL)

The QUAL parameter is the phred-scaled quality score probability of the alternative allele at a given site in sequencing data being wrong. This scale is calculated as:

$$QUAL = -10 * \log (1-p)$$

where p is a base-calling error probability. A value of 10 indicates one in 10 chance of error, while a value of 100 indicates one in 100 chance. Higher QUAL values indicate higher confident in the variant calls. I plotted the QUAL scores for eight different callsets based on filtered and unfiltered variants from Samtools, GATK or both against the Ts/Tv ratio (Figure 2-10). The Ts/Tv ratio was at its highest when variants are called by both GATK and Samtools and pass the callers internal filters (Figure 2-10, dashed red line) and dropped slightly below 3 when the QUAL was < 30, which I used as the minimum accepted threshold.

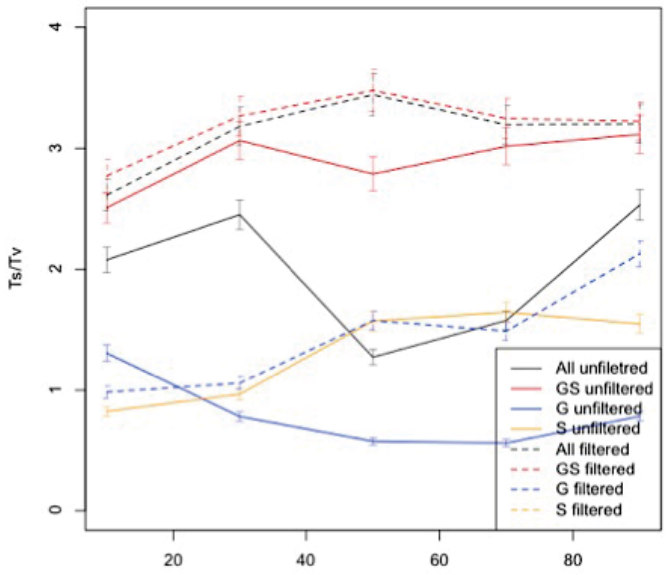


Figure 2-10 The relationship between variant calling quality (QUAL) and the transition/ transversion ration (Ts/Tv) of coding SNVs. The plot shows eight different callsets based on variants called by a single caller or two callers and whether the internal filters of a caller were applied (filtered) or not (unfiltered). These internal filters are usually part of the pipeline itself. (S) is a variant callset called by Samtools alone, (G) variants called by GATK alone, (GS) variants called by both Samtools and GATK, and (All) is a callset composed of variants from the previous three callsets. The GS filtered callset (dashed red line) is the only callset that shows a Ts/Tv ratio close to the expected range (2.9-3.3). However, since the Ts/Tv ratio of this callset drops below QUAL of 30, I used this value as the minimum threshold of high quality variants. Any variants with QUAL < 30 were excluded from the downstream analyses.

Quality by depth (QD)

The QD is a simple statistic to quantify the variant confidence given as 'variant confidence' (from the QUAL field) divided by 'unfiltered depth of non-reference samples' where low QD scores are indicative of false positive calls [275]. QD is only available for variants called by GATK only and thus I was not able to test variants called by Samtools (Figure 2-11). Similar to the QUAL metric above, the variant callset closest to the expected Ts/Tv ratio is the one called by both GATK and Samtools and has passed their internal filters (dashed red line). Unfiltered variants with QD < 5 has significantly lower Ts/Tv ratio below 2.0, which is the minimum accepted threshold I chose for QD (Figure 2-11).

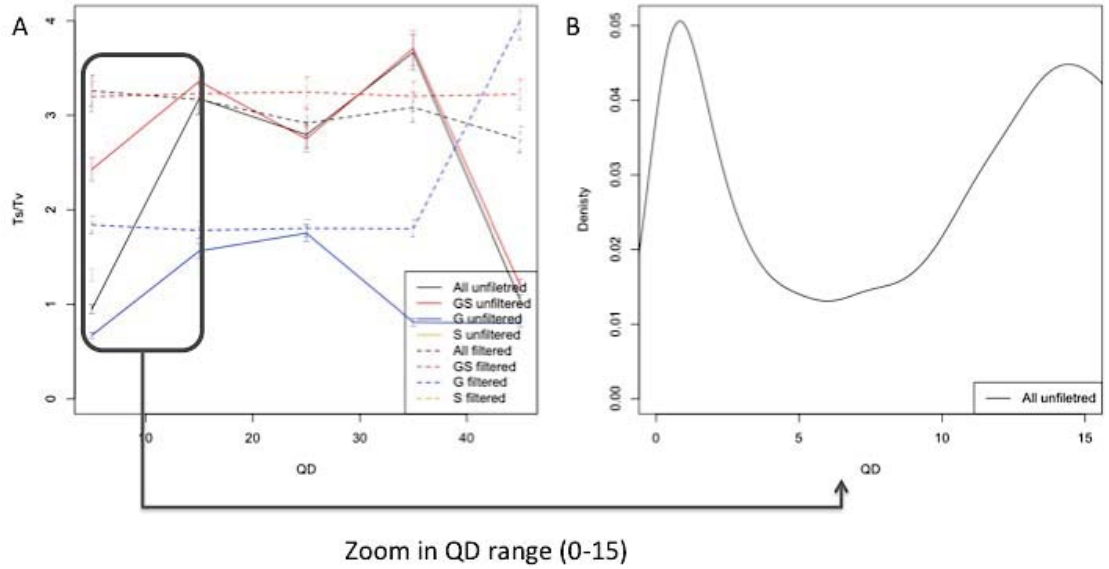


Figure 2-11 The relationship between quality by depth (QD) and the transition/ transversion ration (Ts/Tv) of coding SNVs. (A) I plotted QD values from eight different callsets as described in the previous figure (Figure 2-10). QD values are available for GATK variants, thus variants called by Samtools alone are not shown. The GS filtered callset (dashed red line) the closest Ts/Tv ratio to the expected range (2.9-3.1) is and was consentient along QD values on the X axis. (B) To choose the appropriate minimum QD threshold, I plotted the QD values of all variants, regardless of the caller, from unfiltered callset (All unfiltered, black dashed line in plot A) and restricted the QD to values between 0-15. This shows variants with QD < 5 are enriched for low quality variants (i.e. did not pass the internal filters).

Strand bias (SB)

The third filter I assessed was the strand bias (SB) metric, which quantifies the evidence of a variant being seen on only the forward or only the reverse strand in the sequencing reads. Higher SB values > 0 denote significant strand bias and

are associated with lower values of Ts/Tv ratio, therefore they are more likely to indicate false positive calls (Figure 2-12).

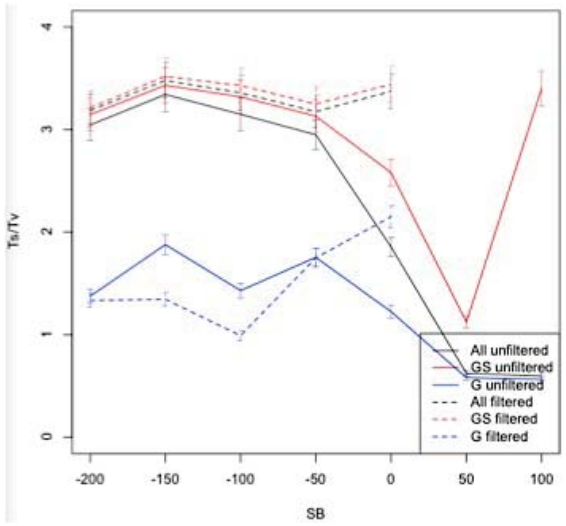


Figure 2-12 The relationship between strand bias (SB) and the transition/ transversion ration (Ts/Tv) of coding SNVs. I plotted SB values from eight different callsets as described in the previous figure (Figure 2-10). At the time, SB values were available for GATK variants only and thus variants called by Samtools are not shown. The callset with closet Ts/Tv ratio to the expected range (2.9-3.1) is the GS filtered callset (dashed red line) and was consentient along SB values (-0.01 to -200). The Ts/Tv ratio values drop dramatically when SB > 0 (solid lines).

Genotype quality (GQ)

Finally, the GQ is another phred-scaled score that represents the confidence of the true genotype at a certain locus. In a diploid genome, the homozygous reference, heterozygous , and homozygous non-reference genotypes are denoted ('0/0' , '0/1' and '1/1') respectively in the variant call format files (VCF files). For a heterozygous genotype (0/1), the genotype quality (GQ) is calculated as :

$$\frac{L(0/1)/L(0/0)}{L(0/1)/L(1/1)}$$

where L is the likelihood of a genotype given the NGS sequence data at that locus. Variants with a GQ of < 30 tend to have lower Ts/Tv ration (~2.7) and hence I used this as the minimum cutoff (Figure 2-13)

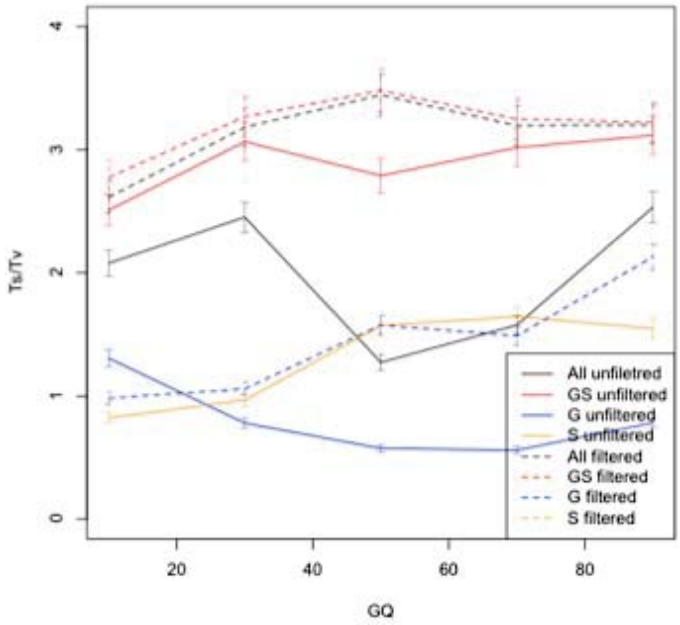


Figure 2-13 The relationship between genotype quality (GQ) and the transition/ transversion ration (Ts/Tv) of coding SNVs. I plotted GQ values from eight different callsets as described in the previous figure (Figure 2-10). The callset with closet Ts/Tv ratio to the expected range (2.9-3.1) is the GS filtered callset (dashed red line) when GQ values > 30.

These four filters where used at the early stages of my analyses of UK10K data to improve the sensitivity and specificity of SNVs calling. It is important, however, to notice that choosing the best filters with highest sensitivity and specificity remains an active area of research. As the developers keep tuning the underlying statistical models in their variant calling programs, these filters need to be adjusted accordingly to reflect the current best practices. More importantly, reviewing the results of validation experiments using capillary sequencing periodically is essential to gain insights about the performance of each filter.

2.3.2.2 Merging caller sets and caller priority

In order to increase the confidence of variant calls, the GAPI pipeline used two independent callers with different underlying probabilistic statistical models to detect SNVs and two callers for INDELs [152-154, 276]. GATK and Samtools were used to call SNVs and while Samtools and Dindel are used to call INDELs. Since Samtools are used to call both SNVs and indels, the GAPI pipeline generates three

files, one from each caller in a variant call format (known as VCF files) [161], per sample.

Using three files separately would complicate downstream analyses since two callers do not agree on the total number of variants, genotypes, and alternative alleles. For example, two SNV callers may detect different alternative alleles at a given locus or report different genotypes (e.g heterozygous by one and homozygous non-reference by the other). To overcome this issue, I decided to merge the three VCF files into a single file per sample. This would have been an easy task if the two callers agreed on all variants, but since this is not the case, I needed to decide on which caller of the two, generated a more reliable set of variants and thus should be used in the conflict cases.

To answer this question, I generated seven different callsets, (Table 2-6 first column) where each callset is composed of at least one group of variants from five scenarios (from 1 to 5). These five scenarios are based on the variant's status according to the two callers (A and B). A variant status can have one of three possible values: (PASS) when a variant is called and passes the caller's filters, (Non-PASS) when a variant is called but does not pass the caller's filters (e.g. when a variant has a low genotype quality), and third status (Not called) is when a variant is missed completely by the caller. Based on the variant status in the two callers, there are five scenarios and each callset is composed of variants from one or more scenarios.

One benefit of organizing variants in these callsets is to test various levels of stringency. For example, the callset named 'Any PASS' includes variants from all five scenarios regardless of the variant status. On the other hand, the callset named "both PASS" includes only variants that pass the called and pass the filters of both callers. These different levels of stringency allowed some callsets to have more variants than other and thus reflected different levels sensitivity and specificity. Moreover, I generated these callsets for both SNVs and INDELs separately (Table 2-7) since SNVs are called by GATK (G) and Samtools (S) while INDELs are called by GATK (G) and Dindel (D).

To decide which callset has the most desirable properties, I measured three different ratios. First, I used the Ts/Tv ratio for the SNVs the expected values ranges between (2.9-3.3) based on different sequencing projects at the Wellcome Trust Sanger Institute and 1000 genomes consortium. For INDELs, I used the coding in-frame/frameshift (n3/nn3) ratio, which was expected to be above 1 where the premise is coding frameshift variants are under much stronger negative selection. The third ratio I used was the rare/common ratio for both SNVs and INDELs (rare variants are defined as MAF < 1%).

Table 2-6 The criteria of choosing different variant callsets in order to determine the closest set to the truth measurements (Ts/Tv, n3/nn3 and rare/common ratios).

Scenarios	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Callset name	Caller A: PASS	Caller A: PASS	Caller A: Non PASS	Caller A: PASS	Caller A: Not Called
	Caller B: PASS	Caller B: Non PASS	Caller B: PASS	Caller B: Not Called	Caller B: PASS
Both PASS	Yes	-	-	-	-
Any PASS	Yes	Yes	Yes	Yes	Yes
Priority PASS (single Caller)	Yes	Yes	-	Yes	-
Any PASS (stringent)	Yes	Yes	Yes	-	-
Priority PASS (stringent)	Yes	Yes	-	-	-
Priority PASS (plus)	Yes	Yes	-	Yes	Yes
No Conflicts	Yes	-	-	Yes	Yes

The total number of SNVs varies between the callsets (Figure 2-14-A). The variation in coding SNVs was observed in the Ts/Tv ratio as well as rare/common ratio (Figure 2-14-B and C). As expected, the most stringent callset (bothPASS), that includes a variant only if it is called by both callers (GATK and SamTools) and passes both of their filters (i.e. PASS), has the highest Ts/Tv ratio (~3.18) while (anyPass) callset has the lowest Ts/Tv ratio (~3.01).

Table 2-7 A list of callsets in each call set based on the caller and if the pass the caller's internal filters (i.e. PASS).

SNVs		INDELs	
Callset Name	**Callset included**	**Callset Name**	**Callset included**
Both PASS	GS	Both PASS	DS
Any PASS	GS, Gs, gS, G., .S	Any PASS	DS, Ds, dS, D., .S
G Priority PASS	GS, Gs, G.	D Priority PASS	DS, Ds, D.
S Priority PASS	GS, gS, .S	S Priority PASS	DS, dS, .S
Any PASS (stringent)	GS, Gs, 'gS'	Any PASS stringent	DS, Ds, 'dS'
G Priority PASS (stringent)	GS, Gs	D Priority PASS (stringent)	DS, Ds
S Priority PASS (stringent)	GS, gS	S Priority PASS (stringent)	DS, dS
G Priority PASS (plus)	GS, Gs, G., .S	D Priority PASS (plus)	DS, Ds, D., .S
S Priority PASS (plus)	GS, gS, G., .S	S Priority PASS (plus)	DS, dS, D., .S
No Conflicts	GS, G., .S	No Conflicts	DS, D., .S

Keys: A single letter denotes each caller. For example "G" denotes GATK, "S" for Samtools and "D" for Dindel. Capital letter means the variant is a PASS (i.e. passed the caller internal filters) and a small letter if does not pass. The "." means the variant was not called by the caller. As an example, the callset named "G Priority PASS (stringent)" under SNVs includes two types of variants (GS) and (Gs). The (GS) is all variants that are called as PASS in both GATK and Samtools while (Gs) includes all variants that are called by GATK as PASS but called as non-PASS by Samtools.

On the other hand, the rare/common ratio of loss-of-function (or functional variant) shows the opposite trend; "bothPass" callset has the lowest rare/common ratio (~0.09) and "anyPass" showed the highest (~0.15). The benefit of using rare/common ratio is that it can tell us if a certain callset is enriched for rare variant more than expected. Since single-sample variant callers are not aware of the variant frequencies (i.e. whether it is common or rare) one would not expect the callers to be biased towards either rare or common variants. However, the variants called by Samtools seem to be enriched for rare variants mainly in three callsets that use Samtools as the dominant caller (S_Priority, S_ PriorityPASSplus and S_PriorityPASSstringent). What is even more interesting is that the Ts/Tv and rare/common ratios are inversely correlated (Figure 2-14-D). The higher Ts/Tv ratio gets, the lower the rare/common ratio becomes. Additionally, this correlation is also seen in other classes of variants such as functional (missense), silent (synonymous) and intronic variant (data not shown).

Similarly for INDELs, I examined different callsets derived from two callers, Dindel and Samtools (Table 2-6 and Table 2-7). The truth measurement I used for INDELs includes coding in-frame/frameshift (n3/nn3) and the rare/common ratios. Not surprisingly, the most stringent callset is "bothPASS" which includes INDELs that are called both callers and pass their internal filters. This callset performs well on both matrices (the n3/3nn ratio is ~1.66 and the rare/common ratio is ~0.10, see Figure 2-15 A-C). Here again, we see inverse correlation between these two ratios as we saw between the Ts/Tv and rare/common in the SNVs (Figure 2-15-D).

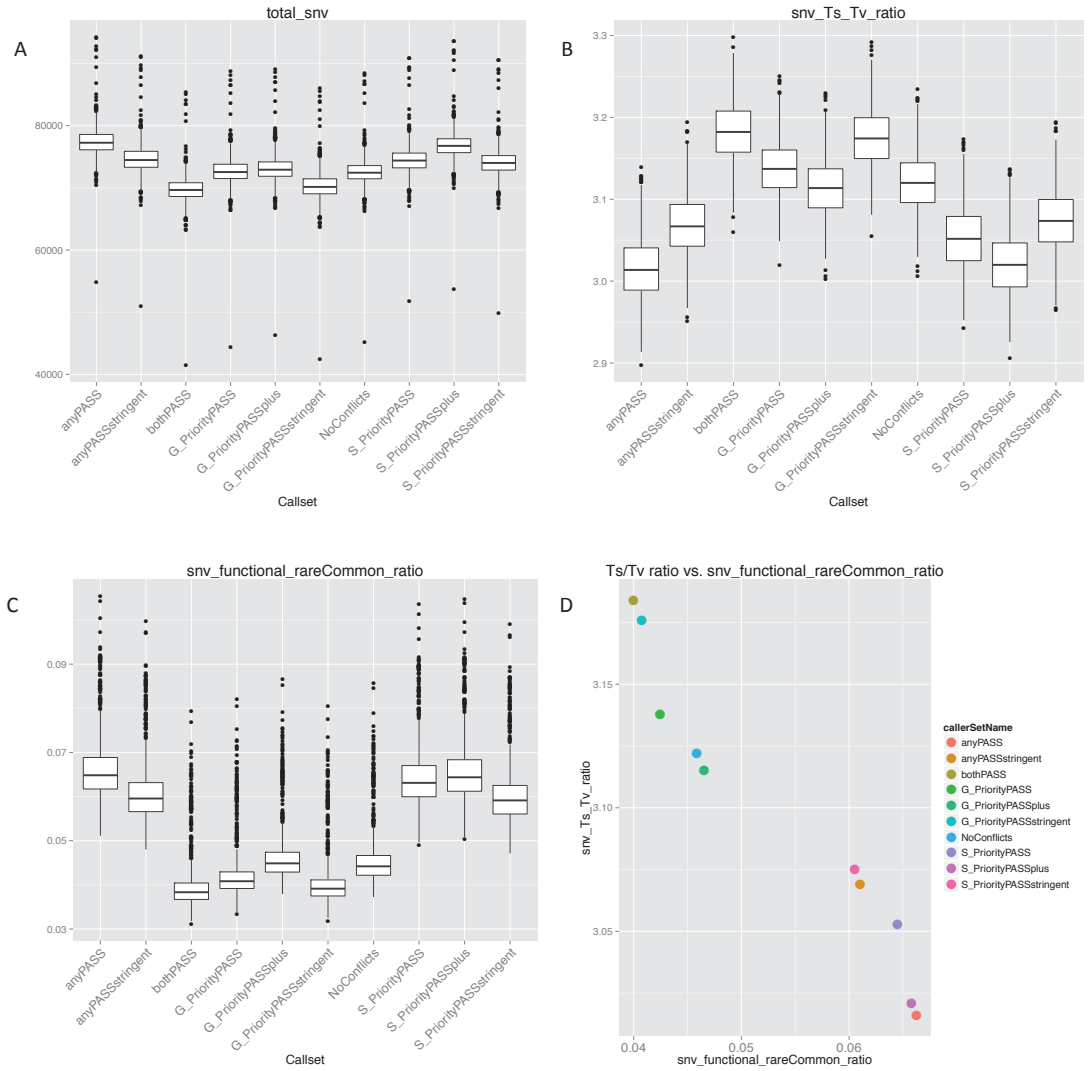


Figure 2-14 **Comparison of SNV callsets from GATK and Samtools.**
(A) Shows the total number of variants in each call set (n=960 samples) and most are comparable. (B) Ts/Tv ratios of functional variants (missense) SNVs per callset. (C) Rare/common ratios of functional variants (missense) SNVs per callset. (D) The relationship between Ts/Tv and rare/common ratios per callset.

Although these analyses were very informative, they were not enough to determine which caller contributed the most to the false positive rate (in terms of low Ts/Tv, n3/nn3 and / or rare/common ratios). The final piece of information was obtained by dissecting each callset to its basic five scenarios as defined in (Table 2-6). For example, SNVs variants can be grouped into five groups (GS, Gs, gS, G. and S.). Similarly, for INDELs, there are five classes (DS, Ds, dS, D. and .S) (see Figure 2-16). This analysis shows that Samtools tends to call more rare variants (in both SNVs and INDELs) and generally performed worse than other callers.

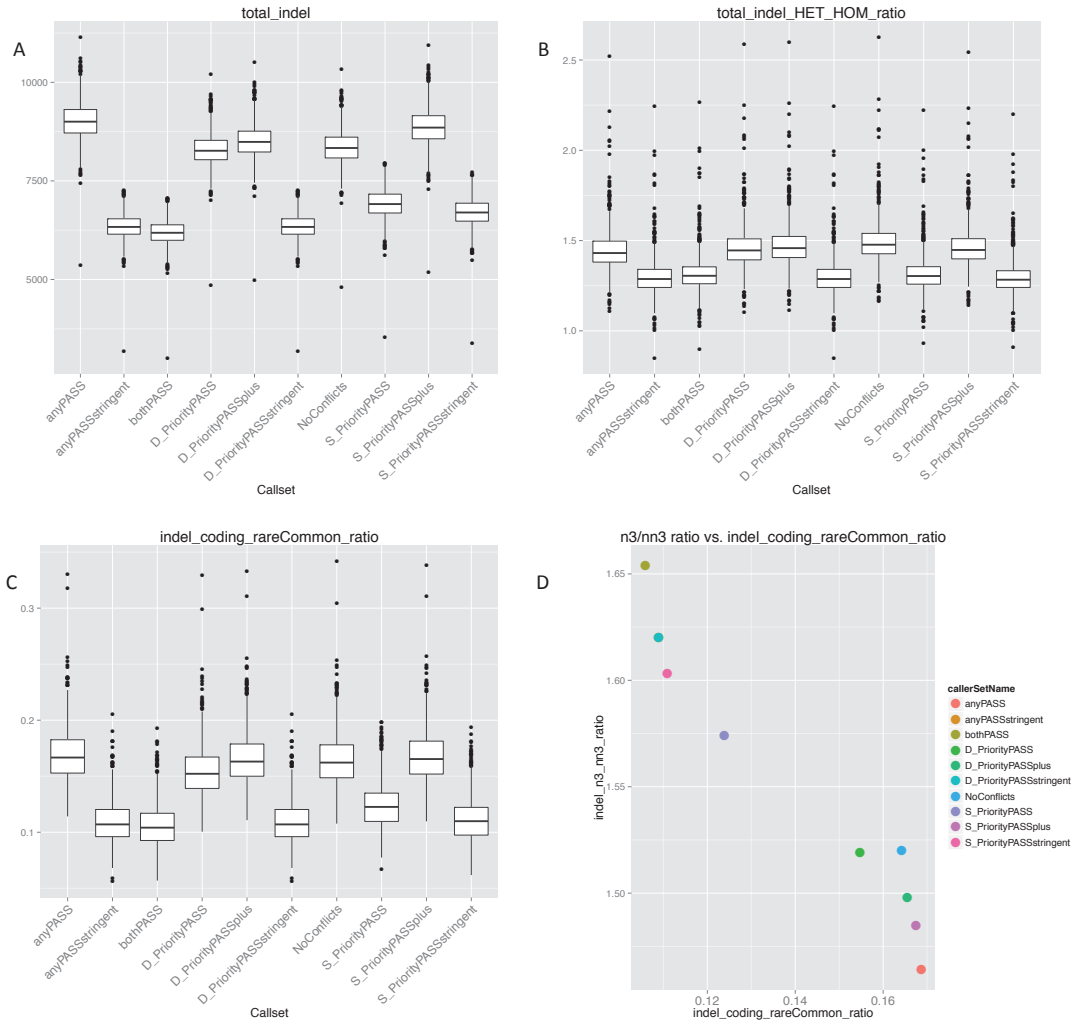


Figure 2-15 Comparison of INDEL callsets from Dindel and Samtools callers.
(A) Shows the total number of variants in each call set (n=960 samples) and most are comparable. (B) In-frame/frameshift (n3/nn3) ratios of coding INDEL variants per callset. (C) Rare/common ratios of coding INDEL variants per callset. (D) The relationship between n3/nn3 and rare/common ratios per callset.

This has a very important consequence on the downstream analysis since, on average, Samtools contributes 2.5 rare loss-of-function SNVs, four rare missense and two rare coding INDELs per sample. These might seem small for the number of candidates in one sample, but in a project with 100 or 1000 samples, this has a tremendous effect on the number of candidate variants needed to be validated or sent for functional studies.

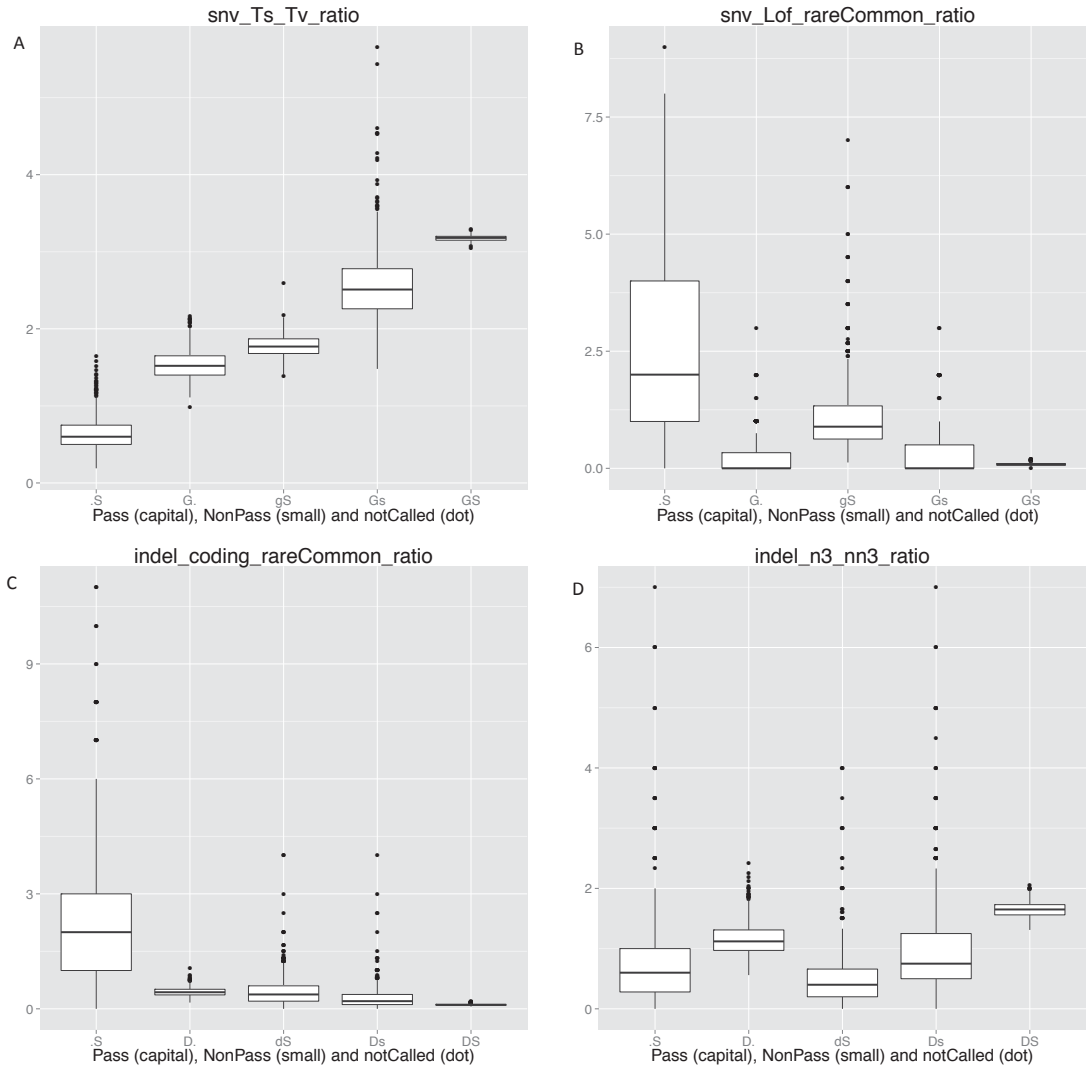


Figure 2-16 **Comparing callsets by callers.**
(A) Ts/Tv ratio of functional (missense) SNVs. (B) Rare/common ratio of loss-of-function SNVs (includes stop gain and variants that disturb the acceptor or donor splice sites). (C) Rare/common ratio of coding INDELs. (D) In-frame/frameshift (n3/nn3) ratio for coding indels. A single letter denotes each caller: "G" denotes GATK, "S" for Samtools and "D" for Dindel. Capital letter means the variant is a PASS (i.e. passed the caller internal filters) and a small letter if does not pass. The "." means the variant was not called by the caller. As an example, the callset named "G Priority PASS (stringent)" under SNVs includes two types of variants (GS) and (Gs). The (GS) is all variants that are called as PASS in both GATK and Samtools while (Gs) includes all variants that are called by GATK as PASS but called as non-PASS by Samtools.

Collectively, these results suggested the importance of discarding or flagging the rare coding variants called by Samtools alone (both SNVs and INDELs) in order to decrease the false positive rare candidate variants. It is important to notice that these observations are true for the specific older version of Samtools and for the filters used in the pipeline and may change accordingly.

2.3.2.3 Sample and data quality control tests

Before obtaining a set of high quality DNA variants for any downstream analysis, several tests are required to detect any quality issues such as contamination, sample swapping or failed sequencing experiments at the level of DNA samples, sequence data (BAM files) and called variants (VCF files).

DNA sample quality tests

The sample logistic 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 tested DNA volume and concentration using PicoGreen assay [277] to make sure every sample met the minimum requirements of exome sequencing. Additionally, 26 autosomal and four sex chromosomes SNPs were genotyped as part of the iPLEX assay from Sequenom (USA). This test helps to determine the gender discrepancies or possible contamination issues. Occasionally, the relatedness between sample and the family membership may need to be tested using the genotype of SNPs in iPLEX assay from the sample sequence data. An example of relatedness test from sequence data is discussed in chapter 3 (part of a replication study of 250 trios with tetralogy of Fallot).

Sequence data quality tests

The second group of quality tests was performed on the sequence reads generated by the next-generation sequencing platform. Carol Scott from the Genome Analysis Production Informatics (GAPI) team performed these tests to detect samples with low sequence coverage.

Variant quality tests

The third group of quality control tests targets the called variants that are stored in the Variant Call Format (VCF) files [161]. The aim of these tests is to detect the 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 (Figure 2-17 for SNV and for Figure 2-18 for INDEL variants). These plots are based on 94 CHD samples generated by GAPI pipeline and these plots are generated for each CHD project in chapter 3 and 4. These serve to monitor the consistency of variant calling between samples from the same project and also between different projects. Samples that show extreme low or high values above 2-3 standard deviations of the mean values are flagged for further investigations to determine the possible causes (e.g. contamination issues, poor sequence data, etc.)

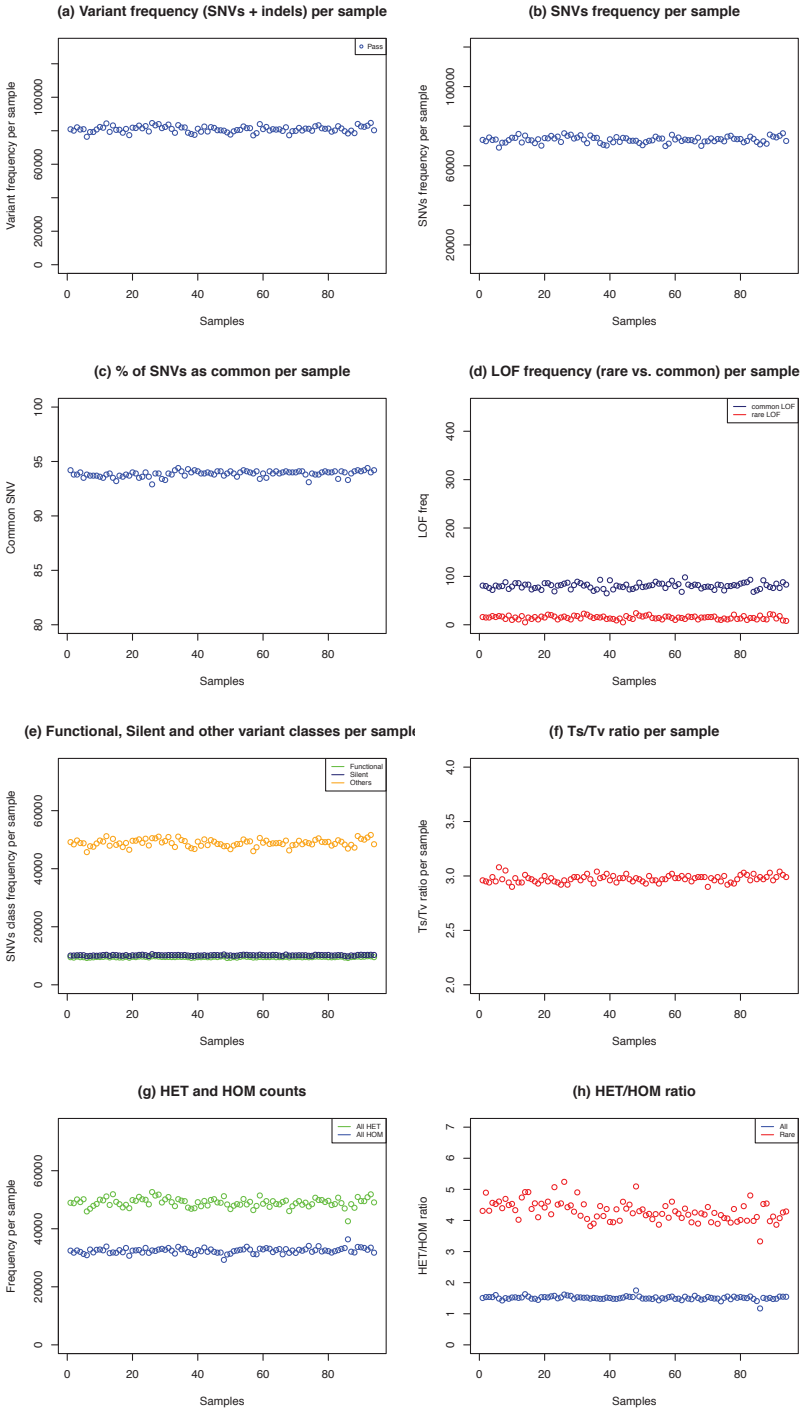


Figure 2-17 An example of QC plots I routinely generate for all samples in each study.
Variant counts per sample (n=94 selected CHD samples). (a) Total number of variants, both SNVs and INDELs, that pass caller internal filters (i.e. PASS). (b) Total number of single nucleotide variants only. (c) Percentage of common variants (MAF ≥ 1% in 1000 genomes project). (d) Number of rare and common loss-of-function (includes stop gain and variants that disturb the acceptor or donor splice sites). (e) Number of functional (missense), silent (synonymous) or others (include non-coding variants such as intronic and variants in untranslated regions, UTR). (f) Transition/transversion ratio of coding SNVs. (g) Count of heterozygous and homozygous variants. (h) Homozygous/heterozygous ratio of all or rare variants.

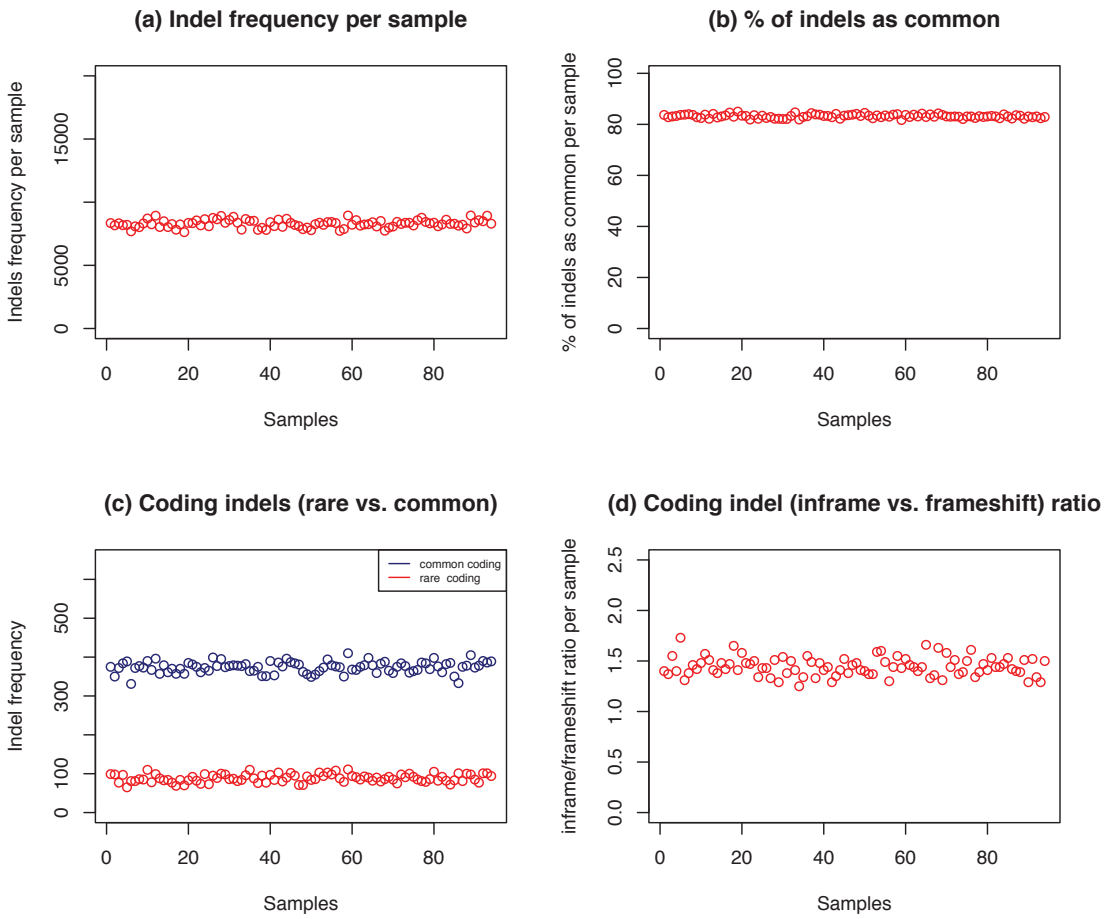


Figure 2-18 Count of INDEL variants per sample (n=94 selected CHD samples). (a) Total number of INDELs that pass caller internal filters (i.e. PASS). (b) Percentage of common variants (MAF ≥ 1% in 1000 genomes project). (c) Number of rare and common INDELs. (d) Coding In-frame/frameshift ratio (n3/nn3).

2.3.3 Minimizing the search space for causal variants

2.3.3.1 Minor allele frequency

In this thesis I have assumed that highly penetrant genetic causes of CHD are rare in the population given the fact that CHD affects usually less than 1% of the population and highly penetrant alleles should be strongly selected against. This makes annotating variants in CHD samples with allele frequency in matching population highly important for downstream analyses such as the family-based co-segregation, case/control and many other analyses. In this section, I describe the different resources of population allele frequencies that I used and their effect on the final number of rare candidate variants.

It is generally accepted that rare variants are defined as the variants with a minor allele frequency of 1% or less [278]. Currently, there are three major projects from which the allele frequency is available in a large number of samples. The first is the 1000 genomes that include 1,092 samples from different populations and used low-depth whole genome sequencing and high-depth whole exome sequencing [155]. The second is the NHLBI Exome Sequencing Project and includes 6,015 individuals of European American and African American ancestry and uses high-depth whole exome sequencing [199]. The third MAF resource is the UK10K cohort of low-depth whole genome sequencing from ~4,000 individuals of European ancestry [264]. While the individuals from the 1000 genomes and UK10K Cohort are presumably healthy, the NHLBI Exome Sequencing Project includes affected patients with various different phenotypes. This led me to disregard the MAF from NHLBI-ESP samples since I cannot rule out the possibility that some samples may have congenital heart defects. Additionally, the captured exome data in NHLBI-ESP project is based on a smaller set of genes (~17,000 genes compared with ~20,000 genes captured in the exome data in my samples), which can adversely affect many downstream analyses such as the case/control analysis by generating spurious false positive signals.

In addition to publicly available MAF resources, I generated an internal MAF based on 576 healthy parents from the Deciphering Developmental Disorders (DDD) project. The main goal of using the internal MAF is to exclude variants that appear as rare according to population MAF recourse but appear in > 1% of the samples. These are expected to be novel 'common' variants or, possibly more likely, sequencing / pipeline errors.

At the time of writing this thesis, there was no general consensus on the best strategy to match the exome sequence variants with variants in population frequency resources, especially the indels, in our internal pipelines (GAPI and UK10K) nor in other external sequencing centers like the Broad institute in the USA (Shane McCarthy, personal communication). Some groups match variants in

their projects with MAF from public resources if both have the same chromosome and position only while others expand this matching strategy by matching variants in a window of 10-30bp to the closest variant.

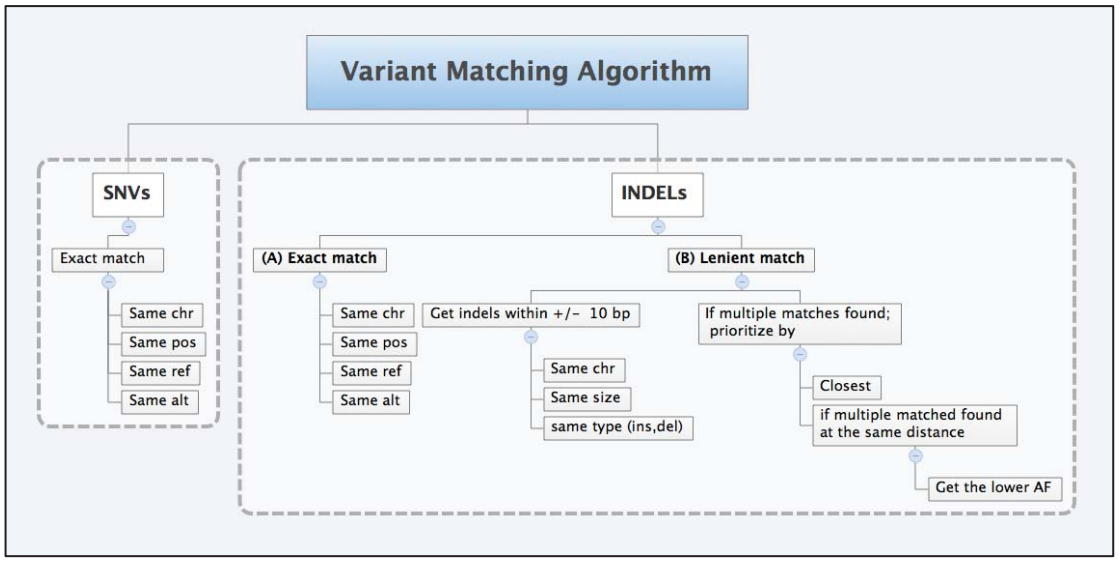


Figure 2-19 The variant matching algorithm between alleles in exome data and alleles from MAF resources.

I designed a hierarchical algorithm that matches between the source files (UK10K, 1KG and ESP) and the target files (CHD samples or other samples like DDD) (see Figure 2-19). The goal of this algorithm is to make sure I match the right allele in my CHD samples with the corresponding alleles in the MAF resources. This algorithm generates two keys; one from the source file (e.g. CHD sample) and the second key is generated from the target file (1000 genomes MAF file) and then tests if both keys match each other (see Figure 2-20 for examples).

In the case of SNVs, I constructed the key using four values (chromosome, position, reference allele and alternative allele) and called this an "exact I" matching. On other hand, INDELs are harder to annotate because callers might call the INDEL alleles differently especially in repeat regions. To accommodate these different scenarios, I tested three different matching definitions. The first is "exact I" which is similar to the SNVs and is considered the most stringent approach. The second strategy is called "exact II" where I construct a key, also using four values (chromosome, position, slice and direction). This key requires

both INDELs in the target and source files to be at the same locus (chromosome and position) while 'slice' is computed based on the DNA sequence difference between the reference and alternative alleles and 'direction' is either deletion or insertion. Although "exact II" matching may look different to "exact I", it is also a stringent matching that tries to accommodate the differences imposed by different callers when they call the same INDEL.

When a matching algorithm fails to find any results using "exact I or II" strategies, it switches to a lenient matching mode where it expands the search for similar INDELs within 10-30bp flanking window. If the algorithm finds more than one INDEL that meet its criteria, it chooses the nearest matching INDEL to the target locus and if it finds multiple INDELs at the same distance, it picks the one with lower MAF value, to be conservative.

(A)

Chr	Pos	Ref	Alt	Direction	Slice
chr1	866511	CCCCT	CCCCTCCCT	ins	CCCT

(B)

Retrieve all variants in 1KG at region chr1: 866511 +/- 10bp flanking region

Chr	Pos	Ref	Alt	Direction	Slice	AF
chr1	866505	G	A	-	-	0.00023
chr1	866511	CCCCT	C	del	CCCT	0.00176
chr1	866511	CCCCT	CCCCTCCCT	ins	CCCT	0.58993
chr1	866517	C	G	-	-	0.00431

Figure 2-20 Example of how MAF matching algorithm works. (A) The chromosome (Chr), position (Pos), reference (Ref) and alternative (Alt) alleles from a source file (e.g. VCF file of a CHD sample). (B) Possible matching alleles within + 10bp flanking region extracted from the MAF resource file from 1000 genomes project. The direction of the allele can be either insertion or deletion in case of INDELs and '-' for SNVs (i.e. point mutation). 'Slice' (red) is the DNA sequence difference between reference and alternative alleles and computed for INDELs only. In this example, since the VCF file contain an INDEL, the matching algorithm will try to look for "exact I" matching key (same chromosome, position, reference and alternative alleles). If this failed, it will start matching using "exact II" strategy (i.e. same chromosome, position, direction and slice), which corresponds to the third record in the (B) where the allele frequency is (0.58993) in the 1000 genomes.

To test the algorithm performance under each mode (exact I, II and lenient), I tested the correlation between three MAF resources (1KG, UK10K and ESP) with DDD internal MAF described above (

Table 2-8). My assumption is that the vast majority of variants should have similar allele frequency in the DDD samples as in the three MAF recourses (except for private or extremely rare variants and sequence errors). A proper matching algorithm should be able to match same alleles and thus the MAF values should show a strong correlation between the DDD samples and the other MAF resources. Both exact I and exact II strategies show a strong correlation between the allele frequencies in 1KG, UK10K or ESP with DDD internal allele frequencies (correlation coefficient > 0.8) but not the lenient strategy for declaring a match (correlation coefficient -0.03 to 0.008).

After I showed that both 'exact I and II' algorithms are well suited for matching alleles in samples sequenced locally with alleles available in public resources, I decided to test the effect of using MAF from different resources on the number of rare coding variants per sample. To evaluate the effect of these MAF recourses, I selected 288 samples from DDD project and annotated them with allele frequency from four MAF recourses (1KG, UK10K, ESP and DDD's internal MAF) (Figure 2-21) in order to eliminate common variants (MAF > 1%). The number of variants left after excluding common variants based on MAF from the 1000 genomes project or the UK10K project was comparable (616 and 631 respectively). The MAF from ESP on the other hand do not appear to be very effective for filtering. This is not unexpected since the ESP sequence data are based on a smaller version of the exome compared with the whole genome data in the 1000 genomes and UK10K projects. However, using all three MAF resources together was more effective than using each separately (~428 rare variants per sample).

Table 2-8 Correlation values between "allele frequencies" of ~9,000 INDELs on chromosome 1 from DDD (n=576 samples) and the corresponding allele frequencies from three population-based projects: 1000 genomes, UK10K twins cohort (n=~4000), and ESP projects (n=~6500) using three matching strategies (exact I, II and lenient). 1KG: 1000 genomes, COHROT: UK10K twins cohort, ESP: NHLBI Exome Sequencing Project, cor=correlation coefficient.

Population-based Projects	Matching strategy		
	Exact type I	Exact type II	Lenient
1KG	0.80	0.83	-0.03
COHROT	0.92	0.73	0.01
ESP	0.89	0.88	0.01

Surprisingly, using the internal MAF from healthy parents in DDD project was even more effective than using all three public MAF together (~419 rare variants per sample when used alone and 327 when used in addition to the other three MAF resources). A possible explanation is that alleles with MAF > 1% and specific to a given project are likely to be sequence or pipeline errors, otherwise they would have been identified in large-scale projects such as the 1000 genomes, which aims to discover alleles with low allele frequency of at least 1% in the populations studied [155].

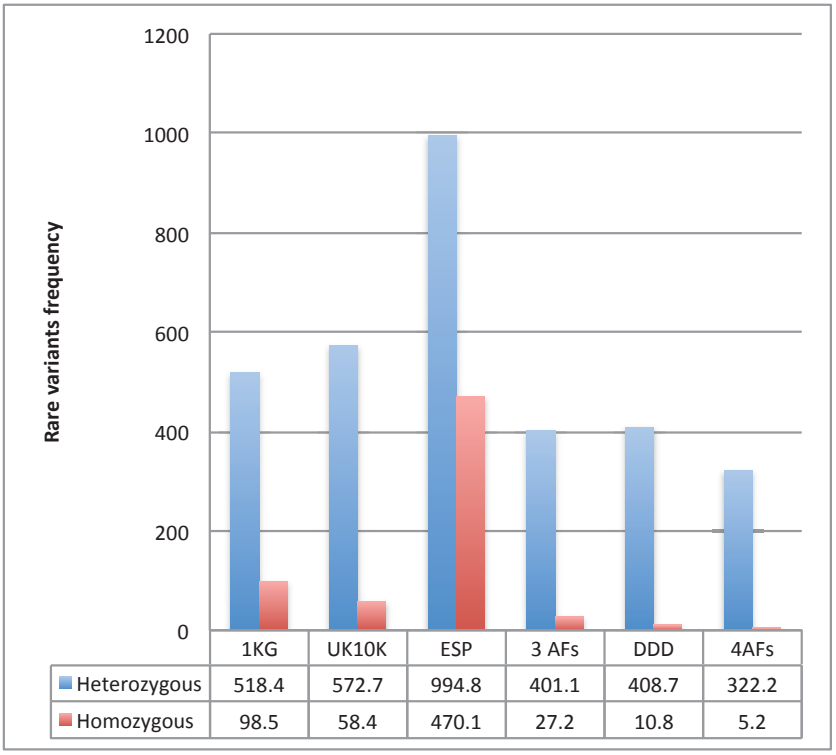


Figure 2-21 average number of autosomal rare variant when filtering based on < 1% minor allele frequencies from different resources. The data are based on 288 samples from the Deciphering Developmental Disorders (DDD) project. 1KG: 1000 genomes, UK10K: 4,000 healthy twins from UK10K cohort, ESP: 6,015 samples from NHLBI Exome Sequencing Project. 3 AFs includes rare

variants in (1KG, UK10K and ESP). DDD is an internal allele frequencies based on 576 healthy parents from DDD project. 4 AFs includes rare variants in 1KG, UK10K, ESP and DDD.

2.3.4 Family-based study designs in CHD

There are many family-based designs one can consider when studying CHD, such as singletons, affected sib-pairs, parent-offspring trios, affected parent-child and multiplex families. However, since the mode of inheritance in CHD is poorly understood in general, there is no obviously optimal study design.

Each design has advantages and disadvantages, for example, in terms of the feasibility of the sample collection and the availability of suitable analytical approaches (Table 2-9). Singletons (or index cases) are the easiest to collect but each sample has several hundreds of rare coding variants if analyzed separately, which makes the task of finding likely pathogenic variants difficult. On the other hand, trio family designs are usually more difficult to collect but they offer a chance to detect *de novo* and definitive compound heterozygous variants in the affected child, which are not feasible in singleton or affected-sib pair designs.

To see how different study designs may affect the final number of candidate genes, I selected one family of healthy parents and three affected children (two females and one male, Figure 2-22) to estimate the number of rare, functional coding variants under different designs and inheritance scenarios. Variants were defined as rare if they have a minor allele frequency < 1% in the 1000 genomes [155] and in 2,172 parents from the Deciphering Developmental Disorders (DDD) project [260] (this analysis was performed more recently with a newer version of the DDD project which include a larger number of parents compared with analysis described in previous sections where I included 576 parents only). Functional coding variants are defined as variants predicted by VEP tool [170] to be either loss of function (stop gain, frameshift or variants affected donor or acceptor splice sites) or functional (missense or stop lost). I excluded silent (synonymous) variants from the analysis.

Table 2-9 Overview of study designs and analytical approaches

Study Design	Advantages	Disadvantages	Analytical approaches
Index cases	-Easy to collect	-Lack of family genotype information means larger search space for causal variant(s).	-Case/control (collapsed, weighted, etc.)
Extended families	- Co-segregated variants that are absent from control provide strong evidence for causality.	- Rare to find and collect samples.	- Linkage analysis and then targeted sequencing.
Trios	- Utilize parental genotype to detected *de novo* variants - Compound heterozygous mutations can be detected - Avoid population stratification bias (e.g. TDT tests)	- More difficult to collect	- *De novo* - Co-segregation - Transmission disequilibrium test (TDT)
Affected-sib pairs	- Suggestive of autosomal recessive disorders. - Small search space due to few autosomal recessive candidates and siblings share only half of the variants.	- The lack of parental genotype information inflates the number of homozygous variant candidates.	- Runs of homozygosity - Co-segregation - Identical By Decent (IBD) analysis (Autozygosity) - Identical by State (IBS) analysis (Allozygous)
Multiplex families **(parents plus > 1 affected child)**	-Combine the power both trios and affected sib-pairs - Smaller search space for variant with more affected children.	- Difficult to analyze when affected members have heterogeneous phenotypes - Less common families than the trios.	Same as trios in addition to the affected sib-pairs
Affected parent-child	- Suggestive of autosomal dominant disorders.	- The variant search space is larger than in trios.	- Co-segregation of heterozygous variants

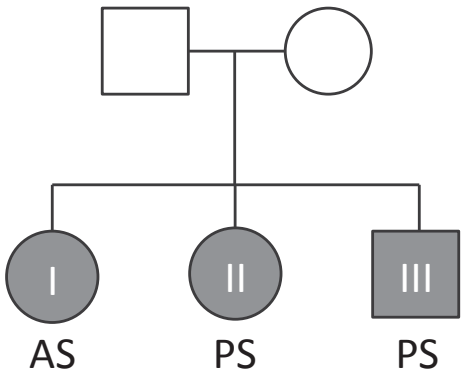


Figure 2-22 Pedigree chart of a multiplex family (three affected children and their healthy parents) used to count the number of candidate genes with rare coding under different inheritance scenarios.
AS: aortic stenosis, PS: pulmonary stenosis.

Initially, I analyzed each affected child separately to test the singleton design and found two rare coding homozygous, 10 compound heterozygous and 381 heterozygous variants on average (Table 2-10). If I consider two children as an affected sib-pair and look for shared rare coding variants, the number of rare coding heterozygous variants drops to less than half and less than a quarter of recessive variants (both homozygous and compound heterozygous) compared with the singleton design. Combining all three affected sibs at the same time shows only 75 rare coding heterozygous variants shared among them, which represents 80% less than singleton and 50% less than two affected sib-pairs but no recessive variants are shared between all three sibs.

On the other hand, the number of candidate genes with rare coding variants drops dramatically to just a handful of genes in the trio design when I consider the parents and assume complete penetrance. This is mainly because the parents' exome data provides additional genotype information to exclude most heterozygous variants (see Table 2-11 for details).

These empirical numbers of rare coding variants shared between different family members are in general agreement with what I would predict from Mendelian inheritance. For example, since the number of rare coding heterozygous variants observed in each child is ~381 on average, two affected sibs should share 50% (IBD=1) or 190 variants which is not far from what I observed in the three affected sib-pairs in this family (~153). Similarly for the rare coding homozygous variants, the observed average in each child is ~10 and each sib-pair is expected to share 25% (IBD=2) or 2.5 homozygous variants, which is very close to the observed value (~2.67).

The variation between the observed and the expected numbers of shared variants under Mendelian inheritance laws is likely caused by under-calling the same variant in one more member. I found the same broad agreement between the average numbers of variants in the affected parent-child pairs (~157) compared with the expected numbers under Mendielian inheritance laws (~190).

Table 2-10 Number of rare coding variants in affected children under different study designs (see family pedigree Figure 2-22).
Singleton: each affected case is analyzed independently. Affected sib-pairs: shared variants between two or more affected sibs without parental information. Trios: each child is analyzed with his/her healthy parents and assuming complete penetrance (see Table 2-11 for the full list of allowed genotypes). Multiple: analysis of two or more children with their healthy parents and assuming complete penetrance.
* Indicates the average number of one affected parent (father or mother) and any child of the three. NA: not applicable (e.g. no autosomal recessive variants are allowed in affected parent-child design).

Family study design	**Samples**	**Number of candidate genes with rare coding variants**		
		Recessive (homozygous)	**Recessive (compound)**	**Dominant**
Singleton	Child I	1	11	373
	Child II	1	12	413
	Child III	4	8	357
Affected sibs	Shared between sibs (I and II)	0	5	162
	Shared between sibs (I and III)	1	1	171
	Shared between sibs (II and III)	0	2	126
	Shared between sibs (I, II and III)	0	0	75
Affected parent-child	One affected parent and one affected child	NA	NA	157*
	One affected parent and two affected children	NA	NA	74*
	One affected parent and three affected children	NA	NA	37*
Trios	Trio (child I)	0	3	1
	Trio (child II)	0	5	0
	Trio (child III)	0	5	0
Multiplex	Shared between trios (I and II)	0	4	0
	Shared between trios (I and III)	0	0	0
	Shared between trios (II and III)	0	4	0
	Shared between trios (I, II, III)	0	0	0

Finally, I consider the shared rare coding variants between two or more trios (i.e. multiplex family design). This study design has identified four genes only with compound heterozygous that are shared between child-I and child-II and another four genes between child-II and child-III. No rare coding variants were detected when all three sibs and their parents were analysed at the same time. This may suggest either a possible under-calling of a monogenic variant (i.e. missed by the callers) or an oligogenic nature of the disease (i.e. multiple genes with different rare causal variants). Nonetheless, the trio design is clearly superior to the affected-sib pairs or singleton designs since it identifies very small number of candidate genes.

Table 2-11 The accepted genotype combinations in a complete trio are the genotypes that are compatible with Mendialin inherence laws and also in agreement with the assumption of complete penetrance. Each trio includes an affected child (male or female) and two healthy parents. Each cell in the first column "genotype combinations" represents three genotypes in child, mother and father. "0" indicates a homozygous reference genotype, "1" is a heterozygous genotype, and "2" is a homozygous genotype in diploid chromosome (autosomal) or hemizygous in a haploid chromosome (e.g. X-chromosome in a male child). Y-chromosome and mitochondrial DNA are omitted from the table. Empty cells indicate that a given genotype combination is incompatible with Mendelian laws (e.g. 1,0,0 is *de novo*) or not expected under complete penetrance assumption (e.g. 1,1,1 is heterozygous in both the affected child and his parents). Only three genotype combinations were considered when I performed trios or multiplex analysis.

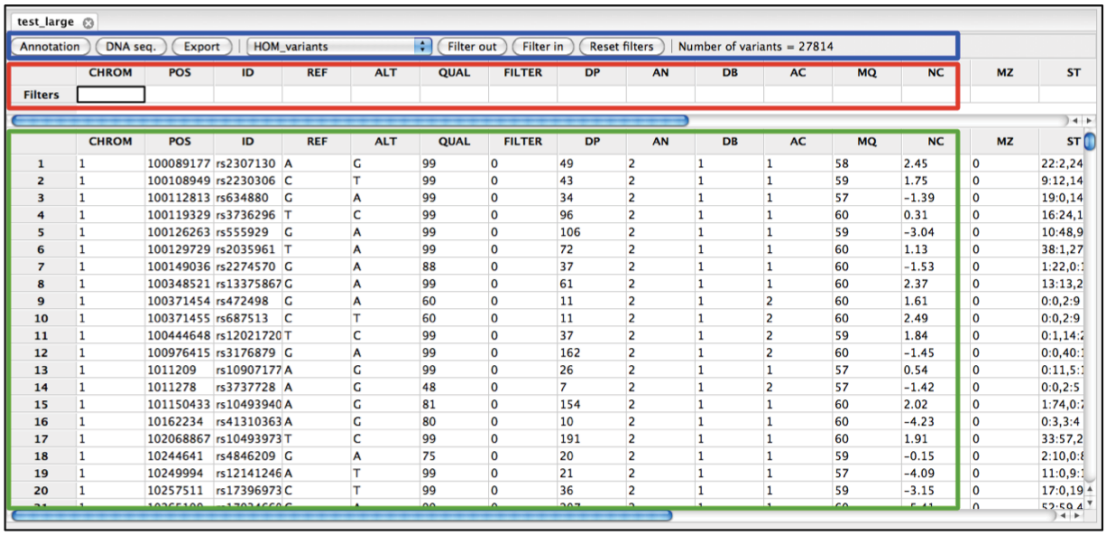
Genotype combinations	**Autosomal**	**X- chromosome in an affected male child**	**X- chromosome in an affected female child**
(1, 0, 0)			
(1, 0, 1)			
(1, 0, 2)			
(1, 1, 0)			
(1, 1, 1)			
(1, 1, 2)			
(1, 2, 0)			
(1, 2, 1)			
(1, 2, 2)			
(2, 0, 0)			
(2, 0, 1)			
(2, 0, 2)			
(2, 1, 0)		Hemizygous inherited from a carrier mother	
(2, 1, 1)	Homozygous in child and inherited from carrier parents		
(2, 1, 2)			
(2, 2, 0)			
(2, 2, 1)			
(2, 2, 2)			
(1,0,1) and (1,1,0)	Compound heterozygous in the child in a given gene		

2.3.5 Family-based Exome Variant Analysis (FEVA) suite

To generate a list of candidate genes from exome data of a given rare, putatively monogenic, disorder, one needs to go through multiple steps that include excluding low quality variants based on various filters, excluding incompatible genotype combinations with either the study design or the plausible inheritance models (see Table 2-11 for an example of incompatible genotypes with a trio design) and filtering common variants (MAF > 1%) as well as non-coding variants since rare coding variants (except silent) are more likely to have a measurable effect on the phenotype. Performing these steps manually in non-specialized software, such as Microsoft Excel, is time consuming and error prone due to the large number of variants. This is clearly not suitable for large-scale projects of hundreds of samples with different family structures.

To automate the analysis and variant reporting under different Mendelian inheritance models I designed a 'Family-based Exome Variant Analysis' tool. FEVA is a suite of tools that enable users to generate a list of candidate genes under various study designs. FEVA offers two interfaces for the end user. The first interface is a Command Line Interface (CLI) suitable for high-throughput analysis, which can be incorporated into automated data analysis pipelines. The second interface is a graphical user interface (GUI) aimed for low-throughput analysis that is easy to use with minimal training (Figure 2-23). I designed the GUI version of FEVA three years ago when many sequencing projects, such as the UK10K RARE project, was just starting at the Wellcome Trust Sanger Institute. At that time, there was no GUI available for our collaborators to explore variants files (VCF files) with ease. I coded most FEVA components in the Python programing language, which I chose for its readability and agility for prototyping. Since Python is a high-level programming language; it can be slow when performing computer intensive tasks (such as parsing large files which are commonly used in the next-generation sequencing era). However, Python is easily extendable by other low-level statically typed, and thus quite fast, programming languages to overcome this limitation. For example, I have used many C and C++ libraries to parse large exome/genome files. Moreover, I used

graphical user interface components, which are written in C++ (QT library) for fast viewing.



	CHROM	POS	ID	REF	ALT	QUAL	FILTER	DP	AN	DB	AC	MQ	NC	MZ	ST
1	1	100089177	rs2307130	A	G	99	0	49	2	1	1	58	2.45	0	22:2,24
2	1	100108949	rs2230306	C	T	99	0	43	2	1	1	59	1.75	0	9:12,14
3	1	100112813	rs634880	G	A	99	0	34	2	1	1	57	-1.39	0	19:0,14
4	1	100119329	rs3736296	T	C	99	0	96	2	1	1	60	0.31	0	16:24,1
5	1	100126263	rs555929	G	A	99	0	106	2	1	1	59	-3.04	0	10:48,9
6	1	100129729	rs2035961	T	A	99	0	72	2	1	1	60	1.13	0	38:1,27
7	1	100149036	rs2274570	G	A	88	0	37	2	1	1	60	-1.53	0	1:22,0:
8	1	100348521	rs13375867	G	A	99	0	61	2	1	1	60	2.37	0	13:13,2
9	1	100371454	rs472498	G	A	60	0	11	2	1	2	60	1.61	0	0:0,2:9
10	1	100371455	rs687513	C	T	60	0	11	2	1	2	60	2.49	0	0:0,2:9
11	1	100444648	rs12021720	T	C	99	0	37	2	1	2	59	1.84	0	0:1,14:2
12	1	100976415	rs3176879	G	A	99	0	162	2	1	2	60	-1.45	0	0:0,40:1
13	1	1011209	rs10907177	A	G	99	0	26	2	1	1	57	0.54	0	0:11,5:1
14	1	1011278	rs3737728	A	G	48	0	7	2	1	2	57	-1.42	0	0:0,2:5
15	1	101150433	rs10493940	A	G	81	0	154	2	1	1	60	2.02	0	1:74,0:2
16	1	10162234	rs41310363	A	G	80	0	10	2	1	1	60	-4.23	0	0:3,3:4
17	1	102068867	rs10493973	T	C	99	0	191	2	1	1	60	1.91	0	33:57,2
18	1	10244641	rs4846209	G	A	75	0	20	2	1	1	59	-0.15	0	2:10,0:8
19	1	10249994	rs12141246	A	T	99	0	21	2	1	1	57	-4.09	0	11:0,9:1
20	1	10257511	rs17396973	C	T	99	0	36	2	1	1	59	-3.15	0	17:0,19

Figure 2-23 **Screen print of FEVA graphical user interface (GUI).**
This simple interface shows three parts. The green rectangle shows a list of variants and their annotations. Each row represents one variant along with its quality scores and biological information such as gene, variant type, effect on protein, etc. The red rectangle is where the user can enter filter conditions to exclude or include rows. The blue rectangle includes additional functions such as applying a set of pre-defined filters or to export a list of candidate variants to other programs.

Although other tools have been published during my work with similar functionality, such as SVA, EVA and VarSift [261-263], none of them were able to fulfill the needs for my projects. One limitation common to these tools it that they are not suitable for both interactive and high-throughput analysis. Additionally, many of them have hard coded filters, and so lack flexibility, or require a certain formatting that is not necessarily compatible with the VCF files generated by the GAPI or UK10K pipelines (see Table 2-12 for comparisons with FEVA).

The family-based analyses in FEVA go through three steps (Figure 2-24): (1) reduce the search space by applying quality and MAF filters (e.g. exclude common variants, low quality, etc.), (2) identify co-segregating variants in family members (e.g. exclude variants in healthy sib or shared variants between affected parent-child), (3) Group the possibly pathogenic variants by the inheritance model (e.g. recessive or dominant).

Table 2-12 Comparison of four freely available graphical user interface applications for genome or exome analysis. N/A: not available.

Features	FEVA	EVA [262]	SVA [261]	VarSifter [263]
Desktop application	Yes	No	Yes	Yes
User custom annotation	Yes	No	Yes	No
Visualization	No	Basic	Advanced	No
Custom filters	Yes	Hard-coded	Hard-coded	Hard-coded
Whole genome	Yes	No	Yes	No
Accepts compressed files	Yes	No	N/A	No
Family Based analysis	Yes	Yes	No	No (Var-MD)
Memory usage (RAM)	Minimal	N/A	Large	N/A
QC statistics	External module	Yes	Yes	No
Has command-line tools	Yes	No	No	No
Input files	VCF	VCF	VCF & bco	VCF
Cross-platform	Yes	N/A	Yes	Yes

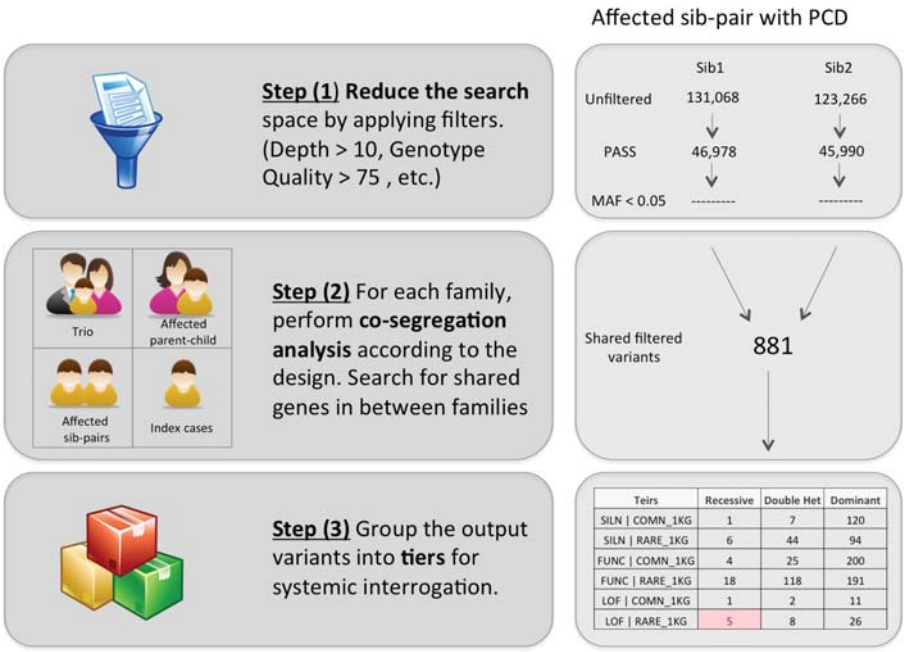


Figure 2-24 **FEVA workflow.**
An example of one sib-pair affected with Primary Ciliary Dyskinesia (PCD), which has been sequenced as part of the ciliopathies study in the UK10K RARE project. The user supplies the variants files and chooses which family design and FEVA performs three tasks automatically. First, FEVA excludes low quality variants and common variants using a MAF threshold supplied by the user. In the next step, FEVA applies the rules of co-segregation designed for affected sib pairs (i.e. shared variants in both sibs). Finally, FEVA groups shared variants under recessive (homozygous or compound heterozygous) and dominant models. Furthermore, FEVA can divide the candidate variants into loss-of-function and functional classes according to the user settings. Almost all steps described here are adjustable by the end user, which enable FEVA to accommodate different needs and scenarios.

The rules of co-segregation vary according to the family design (e.g. singleton, trio of healthy parent or trio of affected father-child, etc.) and can be made more or less stringent. These models are configurable by the user to suit a unique study design (only in the command-line version of FEVA). In the next section, I will describe how I used FEVA with different study designs to identify pathogenic and candidate pathogenic genes for different disorders.

2.3.6 Application of FEVA in rare disease studies

Application 1: Targeted sequencing of linkage regions (monogenic disease)

Dr. Andrew Crosby and his team at St. George's University of London have previously detailed the clinical features of members of a large UK family affected by dominantly transmitted distal hereditary motor neuropathy type VII (OMIM 158580). The team had previously mapped the gene responsible to chromosomal region 2q14 in a family of 14 affected and 12 unaffected members and I collaborated with them to analyze the exome sequence data of one affected family member.

Coding regions were captured with SureSelect All Exons (50 Mb) and sequenced by Illumina HiSeq at the Wellcome Trust Sanger Institute, yielding 9.8 Gb data (~130 million reads) corresponding to 91% target coverage with a mean depth of 1,073 and identifying 52,806 variants. Based on previous linkage analysis [279, 280], I used the FEVA software to report rare coding variants in two regions (~13.5 Mb) with high LOD scores (Table 2-13).

Table 2-13 **Genome coordinates of microsatellite marker**

Regions	Size	Marker ID	Locus in human genome
Region (1)	9.2Mb	AC084377	Chr2:99560750
		D2S160	Chr:2:112998734
Region (2)	4.3Mb	D2S2970	Chr2:118948333
		D2S2969	Ch2:123237183

After filtering common and non-coding variants (Table 2-14), I identified only one loss of function variant within the critical region; this was a single base deletion (c.1497delG) in *SLC5A7* gene encoding the Na^+/Cl^- dependent, high-affinity choline transporter. This novel variant was found to co-segregate in all affected members using capillary sequencing and this work was published in the American Journal of Human Genetics [281].

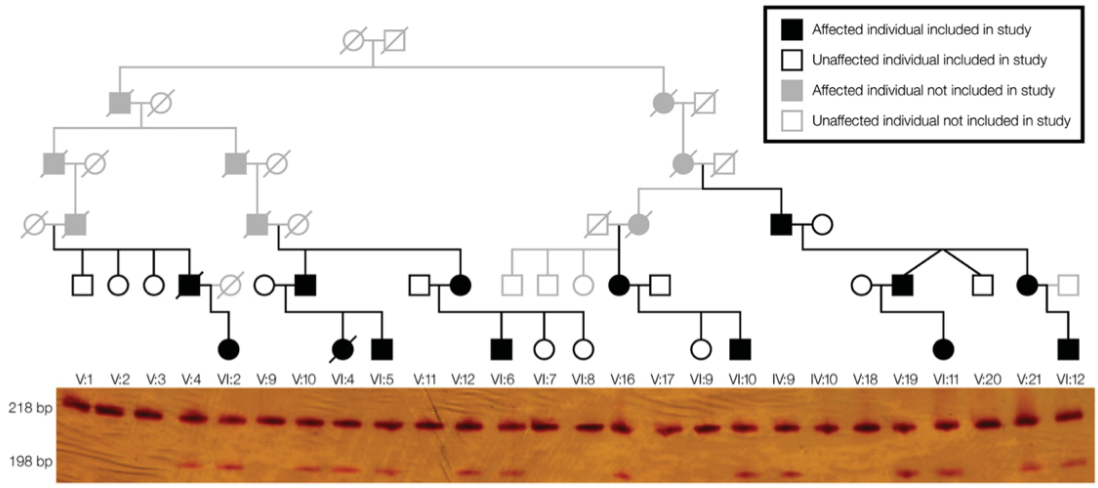


Figure 2-25 Family pedigree and c.1497delG cosegregation in *SLC5A7* gene [281]. The c.1497delG variant results in the creation of a novel *Ssp*I restriction site that facilitates cosegregation analysis by restriction digestion of exon 9 PCR products resolved by polyacrylamide gel electrophoresis. (Image and caption are adapted from [281])

Table 2-14 Number of variants in two linkage regions (~total size of 13.5 Mb). The variants are classified based on genotype (heterozygous or homozygous), by the predicted effect on protein to functional (missense) or loss-of-function (LOF class includes stop gain, frameshift and variants that disturb acceptor or donor splice sites). Only one rare LOF variant, a coding frameshift, found in *SLC5A7* gene that encodes for choline transporter protein.

Genotype	All variants	Common		Rare	
		Functional	LOF	Functional	LOF
Heterozygous	134	32	2	23	1
Homozygous	77	24	1	0	0

Similar to the analytical strategy I used to discover causal mutations in *SLC5A7* gene, I utilized FEVA to analyze data from other monogenic diseases under an autosomal recessive model in collaboration with Dr. Crosby and his team (Table 2-15). In all of these cases, I used the linkage analysis information to guide FEVA while filtering for rare coding homozygous or compound heterozygous variants.

These analyses were usually straightforward since FEVA reported only one or two candidate variants per sample because of the small linkage intervals.

Table 2-15 Results from other monogenic phenotypes where linkage analysis was used to guide the variant filtering of variants from whole exome or custom designed data (using FEVA).

Phenotype	Hereditary spastic paraplegia	Developmental delay with macrocephaly	Microlissencephaly
Mendelian model	Autosomal recessive	Autosomal recessive	Autosomal recessive
Linkage analysis	14.3Mb (chr12)	19q.13.32	2.36Mb (chr19)
Sequencing region	Custom design	Whole exome	Custom design
Number of samples	1	1	1
Candidate gene	*B4GALNT1*	*KPTN*	*WDR62*
Casual variant	c.1458insA	c.776C>T	c.1562T>A and c.4038-4039delAA
Project status	Published in [282]	Published in [283]	Manuscript is being prepared

Application 2: Affected trio families combined with candidate gene screening

The aim of the Deciphering Developmental Disorders (DDD) project is to collect DNA and clinical information from undiagnosed children in the UK with developmental disorders and their parents [260]. I used FEVA to test its performance in high-throughput on 1,080 trios of affected children with various developmental disorders and also to estimate the number of candidate genes, assuming healthy parents and complete penetrance of rare coding variants (Table 2-16).

FEVA was able to report rare coding variants according to the genotype rules in (Table 2-11) under autosomal recessive (homozygous or compound heterozygous) and X-linked models (separately for male and female children). The rare variants are defined as variants with MAF < 1% in the 1000 genomes project and in parental MAF from DDD (n=2,172). Regardless of gender, each child has, on average, four candidate genes with autosomal recessive rare coding

variants (excluding silent) and another three candidate genes on the X chromosome.

I also tested FEVA's ability to screen candidate genes for the presence of rare or novel coding variants (Table 2-16, DDG2P genes). DDG2P is a list of 1,148 manually curated genes with strong evidence supporting involvement in development disorders (the DDG2P gene list was developed by the DDD team). The screening analysis revealed, on average, only one autosomal rare coding variant, one X-linked in females and 0.18 X-linked in males. However, the DDD team implements additional filtering steps for their clinical reporting pipeline. These steps involve matching the phenotype and family history to the genotype (i.e. compatibility with the Mendelian rules), which lowers the number of candidate genes per child still further.

Table 2-16 Number of candidate variants in 1,080 affected DDD trios assuming healthy parents and complete penetrance (558 males and 522 females).
LOF: loss-of-function (include strop gain, variants disturbing acceptor or donor splice sites and frameshift), functional (includes missense). DDG2P: a list of 1,148 manually curated genes with strong evidence supporting involvement in development disorders (the DDG2P gene list is a courtesy of DDD team).

Variant	Chromosome	Genotype	All genes (n=~20,000)		DDG2P genes (n=1,148)	
			LOF	Functional	LOF	Functional
SNVs	Autosomal	Homozygous	0.02	1.01		0.08
		Compound heterozygous	0.13	2.99	0.01	0.42
	X-chromosome (male child)	Homozygous	0.1	3.28	0.02	0.63
INDELs	Autosomal	Homozygous	0.03	0.03		0.08
		Compound heterozygous	0.11	0.07	0.01	0.43
	X-chromosome (male child)	Homozygous	0.07	0.12	0.03	0.66
Total candidate genes in a female child			0.29	4.1	0.02	1.01
Total candidate genes in a male child			0.46	7.5	0.07	2.3

FEVA requires 1-3 minutes to generate a report of candidate genes for one trio. When run in parallel, FEVA can generate reports of candidate genes for thousands of exomes in a few hours with minimum memory usage (< 50 Mb per

trio). This feature makes FEVA suitable for large-scale projects such as the DDD, which aims to analyze the exome data from 12,000 trios in the next couple of years.

Application 3: Affected sib-pairs in UK CHD families

In collaboration with Prof. Eamonn Maher at the University of Birmingham, I analyzed the exome data of 10 families with at least two CHD affected sibs. Two of these families are consanguineous (from Birmingham Pakistani population). All families have two affected sibs except family CHD1 and CHD16 where each has three affected sibs of various CHD phenotypes.

I used FEVA software to generate reports of rare coding variants that are shared between at least two sibs (Table 2-17). The rare variants are defined as variants with MAF < 1% in 1000 genomes and the internal MAF of 2,172 parents from DDD project. As expected, affected sib-pairs from consanguineous families (CHD1 and CHD4) have more candidate genes with autosomal recessive rare coding variants than non-consanguineous families. On average, each family's FEVA output lists 3.5 gene candidate genes with homozygous rare coding variants and 25 candidate genes with compound heterozygous rare coding variants.

Initially, I focused my search for candidate genes with rare loss of function (stop gained, frameshift or variants disturbing acceptor or donor splice sites) (Table 2-18). The top recurrent five genes that appear in most of the families (*ANKRD36C, LINC00955, CDC27, OR4C5,* and *MUC3A*) are unlikely to be linked to the CHD phenotypes since they have compound heterozygous LOF in almost all families. Most of the remaining genes do not have knockout mouse models except three genes (*TTN, PLA2G1B* and *RBMX*) and *TTN* is the only gene that shows structural cardiac defects in the mouse models. Since it not expected to identify recurrent pathogenic genes in such a small study with variable CHD phenotypes, I only considered genes that appear in one affected sib-pair only. I also excluded genes with frameshift variants (INDELs) since they tend to have a

higher false positive rate. Only two genes, *GMFG* and *TAS2R43,* met all filters. TAS2R43 gene encodes a taste receptor and it is unlikely to have a rule in CHD. On the other hand, *GMFG* harbors a rare homozygous stop gain variant (p.Arg24X) in two sibs diagnosed with tetralogy of Fallot in family CHD1 (Figure 2-26). Upon validation with capillary sequencing (carried out by my colleague Chirag Patel), the same homozygous variant co-segregate in the third affected child with TOF (IV:4) but heterozygous in both parents not seen in the fourth child with ventricular septal defect (IV:3) . This variant was absent from ~200 ethnically matched control chromosomes.

Table 2-17 Number of candidate genes with shared coding rare variants, in at least two sibs, under autosomal recessive model.
* Numbers in parenthesis are number of gene candidates with rare coding variants shared between all three sibs.

Family ID	**Consanguineous family**	**Child / Phenotypes**	**Number of sibs**	**Number of candidate genes**	
				Homozygous	Compound heterozygous
CHD1	Yes	Child 1:TOF Child 2:VSD Child 3: VSD, PA (TOF spectrum)	3	23 (1)*	36 (29)*
CHD4		Child 1: VSD, PA (TOF spectrum) Child 2: AS	2	18	24
CHD5	No	Child 1: VSD, RV hypoplasia Child 2: ASD, RV hypoplasia	2	3	21
CHD6		Child 1: TOF Child 2: TOF	2	6	25
CHD11		Child 1: VSD Child 2:AS, BAV	2	1	29
CHD13		Child 1: TGA, VSD, PS Child 2: TGA	2	0	25
CHD16		Child 1: TOF Child 2: VSD, CoA, BAV Child 3: ASD	3	39 (1)*	36 (28)*
CHD20		Child 1: Tricuspid Atresisa Child 2: TGA, RV hypoplasia	2	1	29
CHD22		Child 1: HLHS Child 2: VSD	2	4	19
CHD23		Child 1: AS, subaortic stenosis Child 2: AS, subaortic stenosis	2	0	23

ASD: Atrial Septal Defects, AS: Aortic stenosis, BAV: Bicuspid Aortic Valve, CoA: Coarctation of Aorta, HLHS: Hypoplastic left heart syndrome, PA: Pulmonary Atresia, RV: Right Ventricle, TGA: Transposition of the Great Arteries, TOF: Tetralogy of Fallot, VSD: Ventricular Septal Defects.

GMFG was initially identified as a growth and differentiation factor acting on neurons and glia in vertebrate brain [284]. *GMFG* encodes a small protein of 142 amino acids an actin-binding protein predominantly expressed in microvascular endothelial cells and inflammatory cells [285, 286]. The expression of *GMFG* was found to be unregulated at the site injury during the heart regeneration in

zebrafish models[287]. However, its role in the heart development in mammals has not been studied yet. A knockout mouse of *GMGF* is being modelled at the Wellcome Trust Sanger Institute to investigate further its role during the development of the heart.

Table 2-18 List of candidate genes with rare loss-of-function variants shared in between at least two affected sibs. Genes in red harbor stop gained (SNVs) variants while the rest have frameshift. The phenotypes in knockout mouse models from the Mouse Genome Database [288].

Gene	**Number of families with candidate genes carrying**		**Phenotypes in mouse knockout mouse models Mouse Genome Database**
	Homozygous	Compound Heterozygous	
ANKRD36C		10	NA
LINC00955		10	NA
CDC27		10	NA
OR4C5		9	NA
MUC3A		9	NA
RBMX		5	Decreased lean body mass
CCDC144NL		4	NA
FAM182A		1	NA
TTN		1	First branchial arch and somites, vascular, cardiac and skeletal muscle defects.
MUC4		1	NA
PLA2G1B		1	Abnormalities in lipid absorption and increased insulin sensitivity.
KLHL24		1	NA
ROPN1		1	NA
PITPNC1		1	NA
GMFG	1		NA
TAS2R43	1		NA
ZNF717	1		NA

Next, I performed the same analysis but for shared rare missense variants and identified 119 genes with homozygous and / or compound heterozygous variants in these families (Table 2-19). The majority of the genes appear only in one affected sib-pair while a few appear in all of them (mainly genes from the Olfactory or Mucin gene families which are unlikely to be causal in CHD).

Two of these genes are well known CHD genes such *NOTCH2* and *TBX20* although as dominant genes. Other genes knockout mouse models exhibit structural heart defects (*UTY*, *HSPG2*, *CTBP2*, and *ADAM12*).

Table 2-19 List of candidate genes with rare missense variants shared between at least two sibs. Genes in red have a knockout mouse models that exhibit structural heart defects [288].

Number of Affected sib-pairs	Homozygous	Compound heterozygous
1	*ZC3H13, PGLYRP2, FAM182A, PLCH2, KIAA1683, ZFX, NPIPB1P, PSG6, HR, SHROOM4, PSG11, GMIP, GUCY2F, IKBKG, LPAR4, OR11H6, SPTBN4, UTY, FCGBP, TRGC2, GPKOW, TAS2R43, SLITRK2, MUC16, CXorf61, CXorf64, GPR112, LYNX1, ZNF431, MEGF6, IL12RB1, LRBA, NADK, ZNF30, NKX2-1, ASXL3, OR11H7, MCOLN1, VCX2, OR4L1, TUBGCP5, NDUFA13, HSPG2, TRIT1, OR4K13, PKN2, AQP12A, HNRNPA1L2*	*CTBP2, MYEOV, FILIP1L, FAM182A, TMC2, LRSAM1, CMYA5, KANK1, FAT1, TYRO3, IGHV5-51, MYOCD, TBX20, STIL, SPTBN5, NRCAM, GPR108, MYO15A, PITPNM1, ADAM12, MYO7B, GCOM1, FRAS1, PLA2G1B, LAMB2, RANBP2, IQGAP1, AHRR, PRRC2B, PTGFRN, ODZ4, TRIOBP, HNRNPCL1, KIAA2022, IGHV3-38, NOTCH2, FRG2B, PDHX, AHNAK2*
2	*MUC4*	*FRG1, SRRM2, FAM27E1, USP6, DNAH14*
3	*SLC9B1P1*	*ATM, IGHV7-81, MUC16, ARSD*
4		*PRSS1, CCDC144NL*
5		*TTN, TRGC2, LINC00273*
6		*IGHV2-70, IGLV5-45*
7		*CEP89, NCOR1, RBMX*
8		*TAS2R31*
9		*MUC4*
10		*MUC6, TRBV6-5, ANKRD36C, MUC3A, BCLAF1, OR9G1, CDC27, AQP7, LINC00955, KCNJ12, MUC3A, OR4C5, OR4C3*

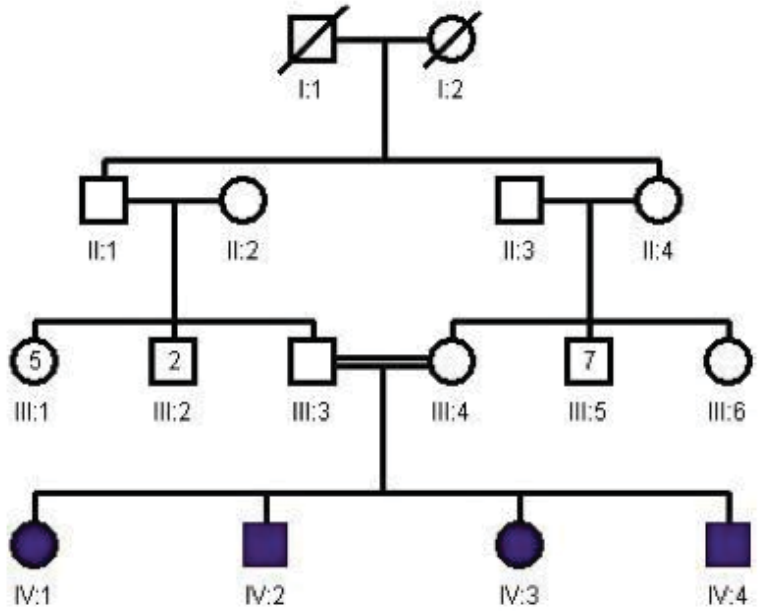


Figure 2-26 **Pedigree chart of family CHD1.**
Four affected sibs from a consanguineous family of a Pakistani origin. Only three sibs had their exome sequenced in this study (IV:1, IV:2 and IV:3). All sibs are diagnosed with tetralogy of Fallot except IV:3 who is diagnosed with ventricle septal defect (VSD). The homozygous stop gain variant was detected in two sibs with TOF (IV:1, IV:2) and capillary sequencing confirmed the presence of the same homozygous stop gain variant in the third sib with TOF (IV:4). Both parents are heterozygous for this variant and in 200 ethnically matched control chromosomes but not see in the child with VSD (IV:3). (Dr. Chirag Patel at the University of Birmingham performed the validation work).

Application 4: Affected parent-child pairs in UK10K CHD families

Most of the samples in UK10K (RARE CHD) are index cases (110 out of 124 samples) except for a few related samples (three affected parent-child pairs, one affected sib-pair and two parent-offspring trios). In this analysis, I focused on the affected parent-child only as this family structure is not covered in the analyses described above. In such a family design, I only looked for inherited rare coding and heterozygous variants shared between the parent and the child.

I used FEVA software to report rare coding heterozygous variants shared between the parent and the child. I defined rare as variants with MAF < 1% in 1000 genomes and the internal MAF of 2,172 parents from DDD project. On average, each affected parent-child pair shared 230 candidate genes (Table 2-20), which is much higher than the number of candidate genes in affected sib-pairs or complete trios (28 and 7 candidate genes, respectively). It is important to note that the number of candidate genes in these families is even larger (47% more) than the number of candidate genes from the simulated parent-child family (see Table 2-10 for details), which has 157 candidate genes on average. This is likely to be as a result in the differences in the calling pipelines (UK10K vs. GAPI). The internal MAF from the 2,172 is based on GAPI pipeline and it is likely to be less effective on samples that went through the UK10K pipeline and thus have more candidate genes per family.

Table 2-20 Number of candidate genes with rare coding heterozygous variants shared between affected parent and child in three CHD families form UK10K RARE CHD project. Loss of function class includes (stop gain, frameshift, variants that disturb acceptor or donor splice sites), functional class includes (missense, in-frame deletion or insertion and stop lost).

Family Id	CHD phenotype		Number of candidate genes	
	Child	Parent	Loss of function	Functional
UK10K_CHD_0015	Atrial septal defect	Atrial septal defect	23	219
UK10K_CHD_0060	Atrioventricular septal defects	Ebstein's anomaly	24	208
UK10K_CHD_0067	Pulmonary stenosis and Atrial septal defect	Pulmonary stenosis	15	201

Since the number of genes with rare functional variants is large in each affected parent-child pair (~200), I focused my search for genes with rare heterozygous loss of function variants (this class includes stop gain, frame-shift, variants that disturb acceptor or donor splice sites) and are shared between the affected parent and the child (Table 2-21). The heart phenotypes observed in these families are varied from family to family and thus I did not expect to see the same gene appear more than once. There are 29 genes where each one has a single loss of function in a single family (first row in Table 2-21). Only one gene, *CCDC39*, shows heart phenotypes in knockout mouse models. This gene harbors a rare frame-shift (c.610_614delTTAGAinsA) in a parent with Ebstein's anomaly and a child with atrioventricular septal defect (family id: UK10K_CHD_0060).

CCDC39 gene encodes a protein that localizes to ciliary axonemes and is essential for the assembly of inner dynein arms and the dynein regulatory complex [289]. Recessive loss of function variants have been found to cause a large proportion of primary ciliary dyskinesia in human. However, the knockdown of Ccdc39 in zebrafish embryos at the 2-cell stage caused a dose-dependent increase in heart looping defects and other laterality defects may suggest a possible *CCDC39* haploinsufficiency [289]. Moreover, a knockout mouse model submitted to the Mouse Genome Database (MGI:5445973) [288] shows double outlet right ventricle, atrial septal defect and dextrocardia but it has not been published. These findings suggest the involvement in *CCDC39* in the development of the heart but further work is required to confirm the role of this heterozygous frame-shift variant in causing the heart phenotypes observed in this family.

Table 2-21 List of genes with rare loss of function (stop gain, frameshift, variants that disturb acceptor or donor splice sites) variants shared between affected parent and child.

Number of affected parent-child pairs	**Genes**
1	*ATXN3L, AXDND1, CCDC39, CCDC7, CCL8, CD5L, COL6A5, CYP2C8, AC061992.1, ERAP1, F5, FAM49A, FHAD1, FLG2, GPLD1, MUC19, NDUFA10, NLRP5, OR51E1, OR51T1, OR5AN1, POLR1A, SERGEF, SMYD4, TAS1R3, TAS2R43, VNN2, VPS8, ZNF211*
2	*PRSS3, RBMX*
3	*CDC27, LINC00955, FRG1B, MUC3A, OR4C5*

2.4 Discussion

NGS has accelerated gene discovery in rare monogenic disorders in the last few years. More than 180 novel genes have been identified using whole genome or whole exome sequence data generated by NGS platforms so far. Based on the current rate of novel gene discovery, it has been estimated recently that most of the disease-causing genes of rare monogenic diseases will be identified by the year 2020 [202].

The success of NGS with rare monogenic disorders inspired me to apply the exome sequencing strategy for studying congenital heart defects (CHD). However, applying NGS to CHD cases is not straightforward since the inheritance model for CHD is not well defined. Evidence from genetic epidemiology and genome-wide association studies has supported the polygenic model [112, 115] and at the same time several monogenic examples of isolated and familial forms of CHD have been reported in the literature [14]. There is no general consensus on what is the most plausible inheritance model that can explain CHD. For this reason, I explored four different family-based study designs in order to evaluate the power of each design to identify rare coding variants that might explain the monogenic CHD cases.

This chapter describes the tools and pipelines used to call single nucleotide (SNVs) and insertion/deletion (INDELs) variants from exome data. One major challenge I addressed is how to improve the sensitivity and specificity of variant calling from exome data. The issue of sensitivity and specificity stem from the underlying probabilistic statistical models implemented by different variant callers. These models are being actively developed and thus it is expected that the best practices for filtering and cleaning up exome data will keep changing for the foreseeable future, especially for indels.

In this thesis, two pipelines have been used to call variants from exome data: GAPI and UK10K pipelines. Both of these pipelines use different callers and

filters to generate the variants. Although they have been able to detect a relatively comparable number of coding SNVs, the number and type of INDELs varied substantially in both pipelines. This is most likely caused by the use of an additional caller, Dindel, to detect INDELs in the GAPI pipeline. On the other hand, the intra-pipeline comparisons between GAPI sample releases at different time points show minimal differences. These findings highlight the need to use only one pipeline for consistency and to avoid unnecessary complications for the downstream analysis (such as case/control analysis using the samples from different pipelines as discussed in chapter 4).

To improve the sensitivity and specificity of SNV calls generated by UK10K pipeline as an example, I tested the relationship between strand bias (SB), quality by depth (QD), genotype quality (GQ) and variant quality (QUAL) with transition/transversion ratio (Ts/Tv) to chose the proper filtering thresholds. Applying these filters has helped me to eliminate low quality variant calls in a systematic fashion. However, this method of variant filtering using hard cut-offs is no longer considered the best practice and newer filters based on sophisticated statistical models that integrate several quality metrics simultaneously have now been used. One example is the Variant Quality Score Recalibration (VQSR) scores recently implemented in GATK, which seems to be superior to other filtering methods. However, VQSR is not so successful for filtering indel callsets since it is sutable for SNV callsets only.

It is not uncommon to use more than one variant caller to detected SNVs and / or INDELs to improve the sensitivity and specificity of variant calling. Theoretically, callers that utilize different probabilistic models to call variants independently, are most appropriate. However, it was not clear how to resolve conflicts that arise when a variant passes the filters of one caller but not the other, or when a variant is missed by one of them. My analysis of 14 different datasets (seven INDELs and seven SNVs) based on different scenarios shows that INDELs called by Dindel were superior to Samtools calls, as they show in-frame/frameshift (n3/nn3) ratio closer to the exacted ~1.5 ratio. Similarly, GATK SNVs calls were superior to Samtools calls in terms of transition /transversion (Ts/Tv) and

rare/common ratios. These results have led me to change the order of caller when I merge calls in the final variant call format files (i.e. I used Dindel as the default caller for INDELs and GATK as the default caller for SNVs). Such a small decision has a large effect on the final number of rare coding variants. For example, Samtools calls more rare loss of function variants than GATK or Dindel. Such that, in large-scale projects, this could mean hundreds of false positive candidate variants that would slow down any downstream analysis or functional studies.

Once an optimal callset of variants is obtained, it is important to exclude common variants based on **minor allele frequencies** (MAF) to minimize the number of candidate variants. There are many population-based MAF recourses available to facilitate this step such as 1000 genomes (1KG), UK10K Twins cohort (UK10K) and the NHLBI Exome Sequencing Project (ESP). Additionally, I generated a fourth MAF resources (called internal DDD MAF) based on 2,172 parental samples generated by GAPI pipeline to target variants that appear as rare variants in the public MAF recourse but are common in the internal samples which likely indicate that they are sequence or calling errors.

Matching alleles between sequenced samples (e.g. DDD or CHD samples) and the population variation resources (e.g. 1000 genomes project) in order to obtain the correct minor allele frequency is straightforward for SNVs but more difficult for INDELs since they can be called differently due to the genomic context such as homopolymer runs for example. To assign the correct MAF, I tested three allele-matching strategies (two exact matching algorithms and one lenient algorithm based on 10-30bp matching window) and I used the correlation between the observed minor allele frequency in DDD samples and the population allele frequency from all three MAF population resources as a metric to compare different matching strategies. I showed that the exact strategies have a stronger correlation between the observed minor allele frequency from DDD samples and population allele frequency from all three MAF population resources.

Using the **exact matching algorithm**, I evaluated the consequence of applying each MAF resource independently and combined on the final number of rare candidate variants in 288 affected samples from the DDD project. This analysis showed that the internal frequency from the DDD project alone was able to eliminate most common variants compared with other combined public MAF recourses. Combining two or more MAF is more effective than using each individually. However, using allele frequencies from ESP and UK10K has some drawbacks. First, ESP includes many affected samples with unpublished phenotype, which may include CHDs and thus cannot be used as controls. Moreover, the targeted exome in ESP is smaller than the exome design used to sequence CHD samples in my thesis, (~16,000 genes and ~20,000 genes, respectively. Similarly, the MAF from the UK10K Twin cohort does not include variants on X-chromosome. For these reasons, I decided on a MAF filtering strategy using the 1000 genomes project data combined with the internal allele frequencies from healthy parents in DDD project to exclude common variants and pipeline errors.

Another factor that affects the final number of candidate variants/genes is the **family design**. I performed a simulation analysis using one multiplex family of three affected sibs and two parents and showed how the number of candidate variants varied between singletons, sib-pairs, parent-child, and complete trios study designs within the same family.

The **Singleton** study design generates the largest number of candidate variants per sample compared with other family-based study designs, unless it is combined with linkage analysis to limit the search in a smaller region. The example of 'distal hereditary motor neuropathies type VII' with two small linkage regions (9.2 and 4.3 Mb) has identified only one candidate gene, *SLC5A7*. This example, in addition to another three genes identified using the same strategy (*B4GALNT1*, *KPTN* and *WDR62*), indicates that finding causal genes by combining NGS and linkage analysis can be powerful and relatively straightforward. Without linkage analysis, the number of candidate genes per sample is usually large especially for dominant disorders. In the absence of

linkage analysis information, sequencing multiple unrelated cases may help to identify the causal gene in monogenic disorder, but can be challenging for extremely genetically heterogeneous disorders such as intellectual disabilities and CHD. In such disorders, a case/control analysis might be more suitable but requires a large number of samples.

The affected sib-pairs design is helpful when looking for shared homozygous or compound heterozygous candidate genes in non-consanguineous families or homozygous candidates in consanguineous families. This analysis has highlighted variants in a few known CHD genes such as *NOTCH2* and *TBX20,* but these genes are mostly known to cause CHD under a dominant model while they have been reported here to harbor rare and presumably recessive variants. It remains to be seen if these variants are pathogenic. Additionally,_I identified novel genes such as *GMFG* with a homozygous stop gain shared between three affected sibs in the same consanguineous family of a Pakistani origin. These candidate genes were found in a single sib-pair only and thus require additional families sharing the same candidate genes to be identified and / or to be confirmed by functional studies. Nonetheless, the number of recessive candidate genes in this design is manageable and provides a chance to investigate the recessive model in different CHD subtypes.

The **trio and multiplex designs** identify far fewer candidate genes than the other designs because of the additional information from the parents. Assuming healthy parents and complete penetrance, each trio has, on average, seven rare inherited coding variants and a smaller number in multiplex families. The small number of candidate genes per trio makes most downstream analyses amenable to further investigations either *in silico* or by functional experiments (e.g. modeling in zebrafish). The design is also suitable for *de novo* analysis, as I will discuss in the next two chapters.

Many of the steps described above are time consuming and error prone when performed manually in non-specialized software such as Microsoft Excel. I designed the **"Family-based Exome Variant Analysis" (FEVA)** tools to

automate applying various quality filters and to report candidate genes from different study designs. FEVA reports candidate variants under different models of inheritance and can be customized by the end users to accommodate new family designs not covered by the program default settings. I used FEVA successfully to find causal genes in monogenic disorders from single cases such as the *SLC5A7* gene in distal hereditary motor neuropathy (type VII) [281] and another three genes (*B4GALNT1, KPTN* and *WDR62*) in various neurodevelopmental disorders (manuscripts were submitted or are being prepared). Other groups at the Wellcome Trust Sanger Institute as well as external groups from Cambridge University, University College London and other institutes, working with different rare disorders such as ciliopathies [290-292], neuromuscular, thyroid disorders and familial hyperlipidemia, have used FEVA to identify mutations in novel or known genes. Moreover, FEVA is also being used in large-scale projects with hundreds of families, such as in the Deciphering Developmental Disorders (DDD) project [260].

The results from this chapter show that at every step of the analysis pipeline small, seemingly insignificant, changes can have a big impact on the numbers of candidate variants being explored. Planning an upgrade of a pipeline, implementing a new version of a caller, modifying a filter threshold are some of the decisions that should not be taken lightly without careful consideration of how such a decision would affect the output. This is especially true in clinical settings where maximum levels of sensitivity and specificity are required for a definitive diagnosis.

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].

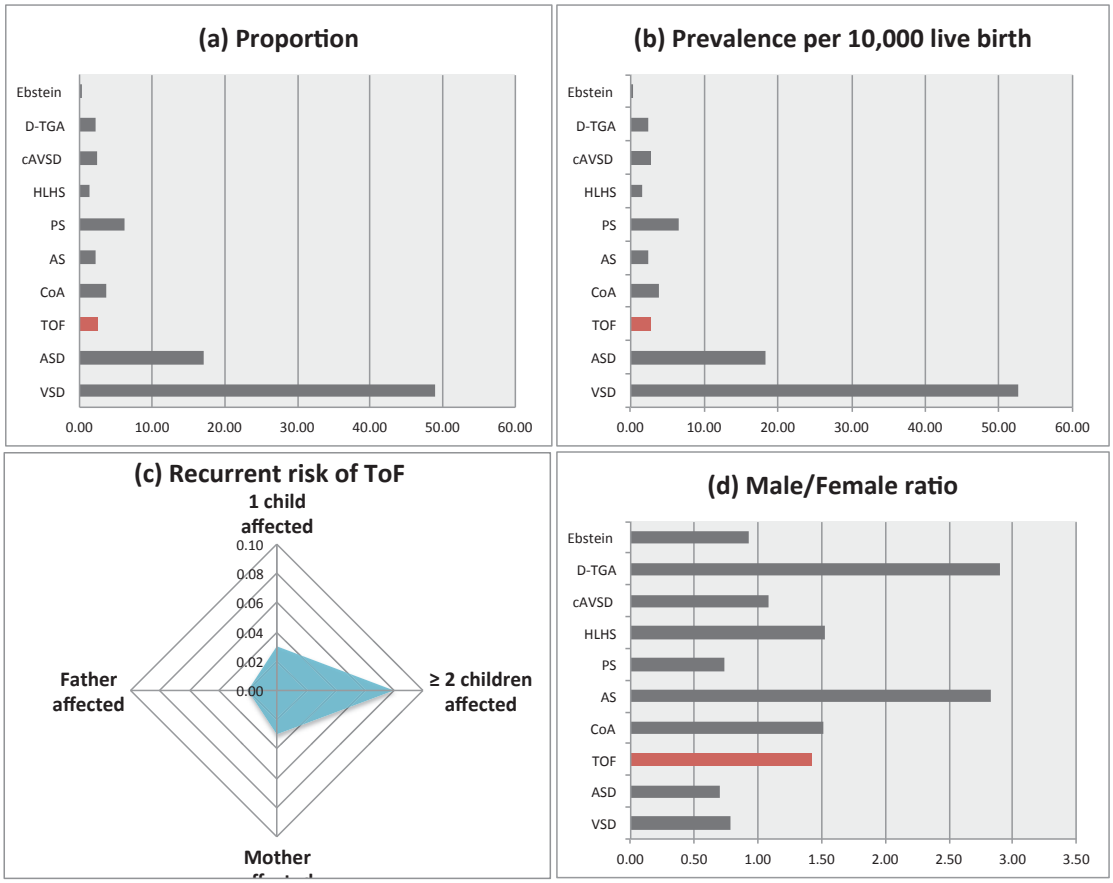


Figure 3-1 (a) Proportion of different CHD subtypes, including Tetralogy of Fallot (red bar) in the PAN registery (n=7,245) during one year 2006-2007 [64]. (b) the prevalence of ToF per 10,000 live brirths from the PAN registry (red bar) compared other CHD cases. (c) Recrrent risk of ToF in first degree-realtives (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].

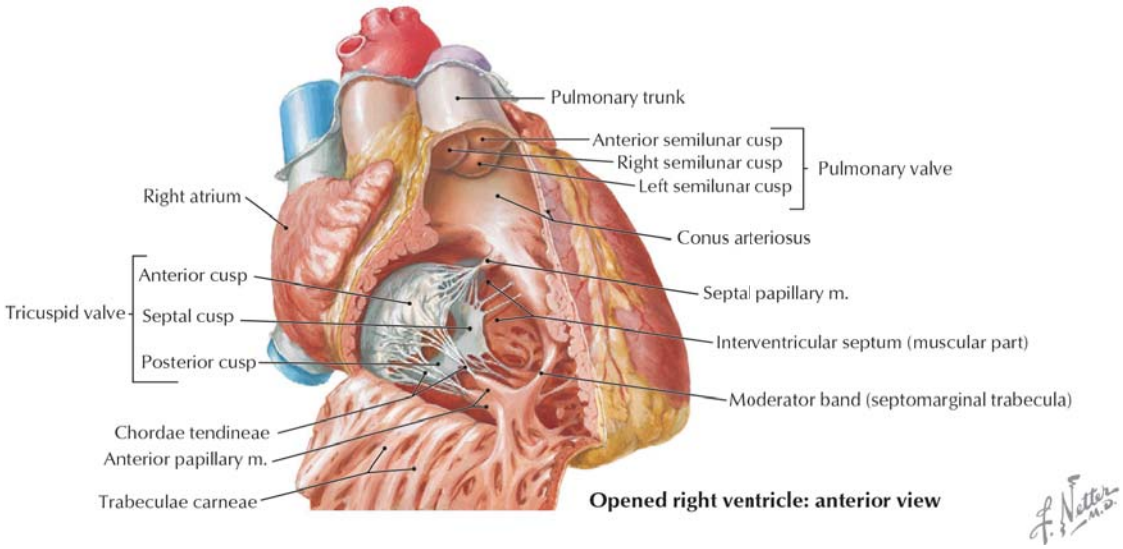


Figure 3-2 The anatomy of the human right ventricle (image adapted from Netter's clinical anatomy [303])

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.

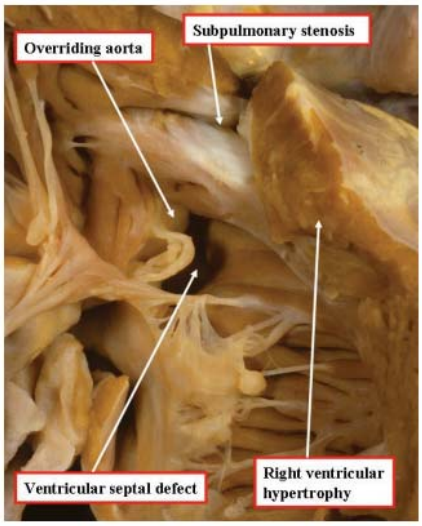


Figure 3-3 The main anatomical features in tetralogy of Fallot (image adapted from [304])

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].

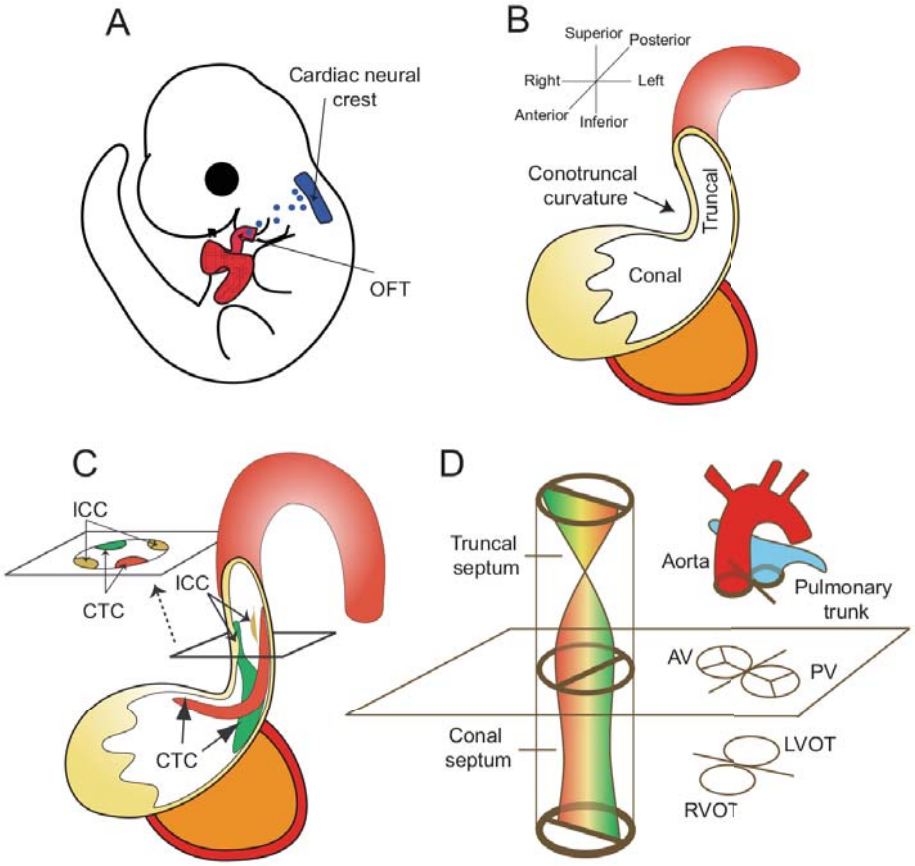


Figure 3-4 Septation of the cardiac outflow tract. (A) Left lateral view of an E10 mouse embryo. The neural crest gives rise to cells (blue) that migrate to and colonize the distal cardiac outflow tract (OFT). (B) The cardiac OFT contains conal (proximal) and truncal (distal) cushions. The boundary between 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.

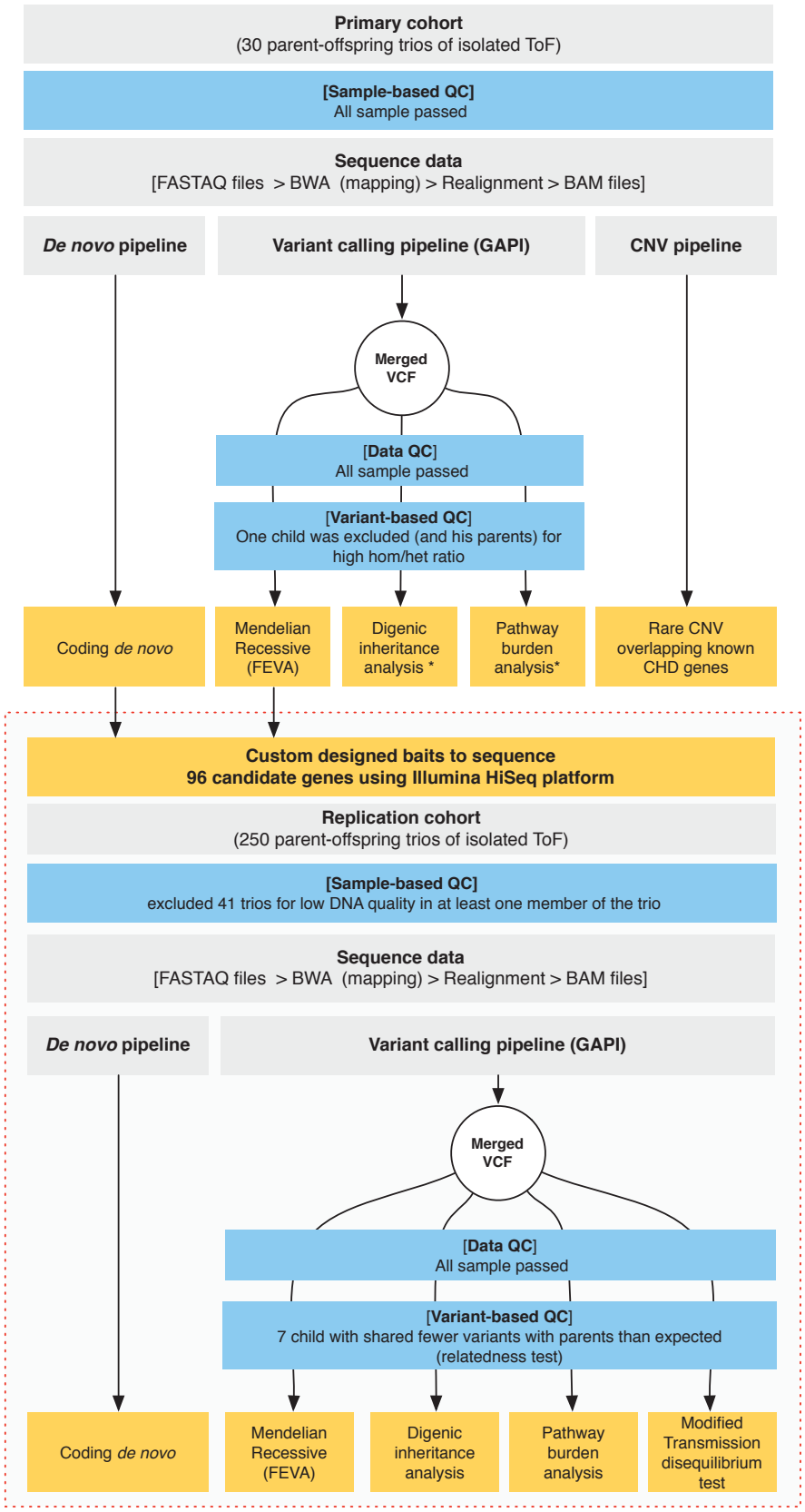


Figure 3-5 A two-stage study design was adapted in this chapter. The first stage included exome sequencing of 30 parent-offspring trios isolated ToF while the second stage included an additional 250 trios as a replication cohort (red dashed box). Quality control (QC) tests (blue boxes) helped to exclude 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

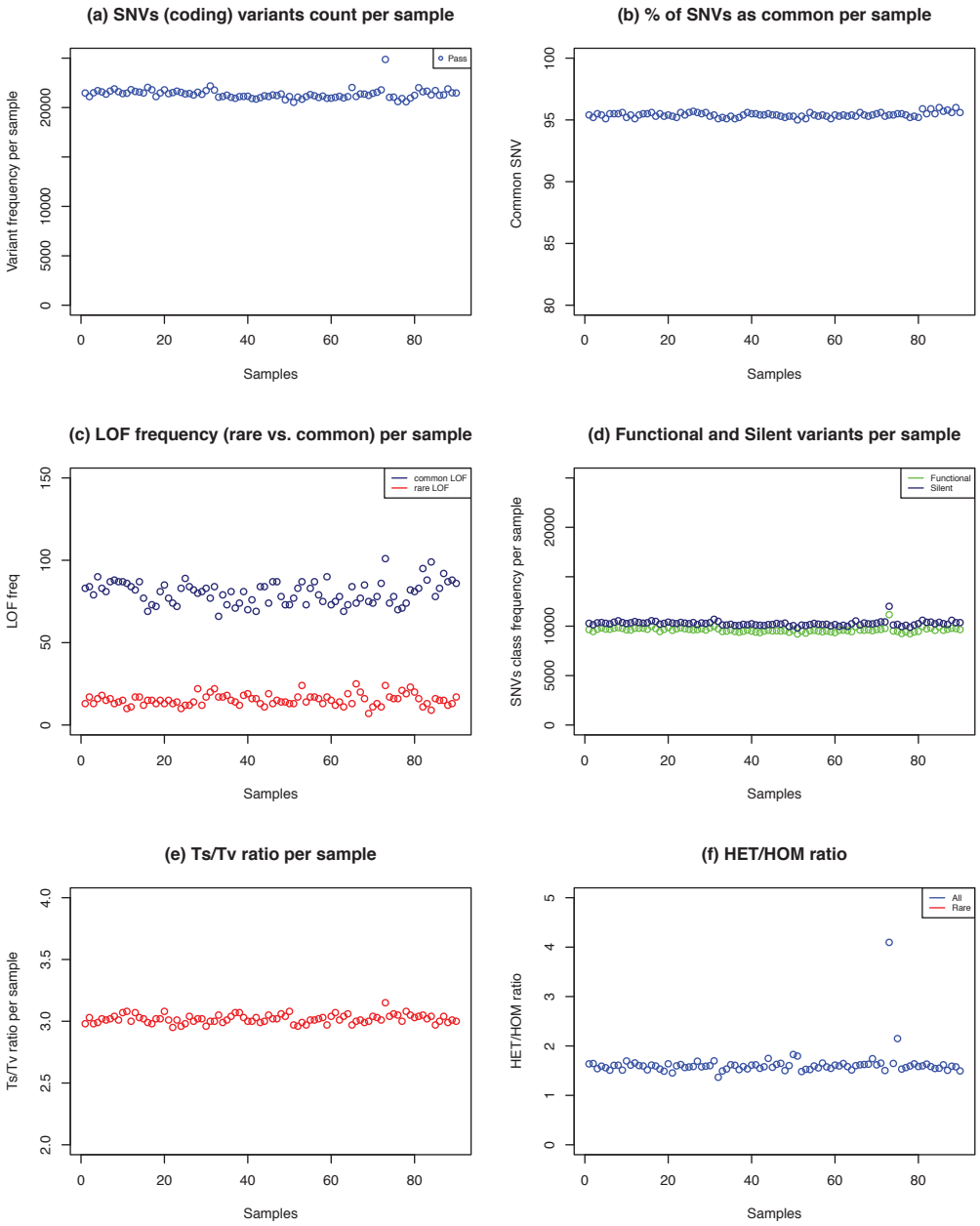


Figure 3-6 Quality control plots including global counts and various single nucleotide variants stats (see main text for description)

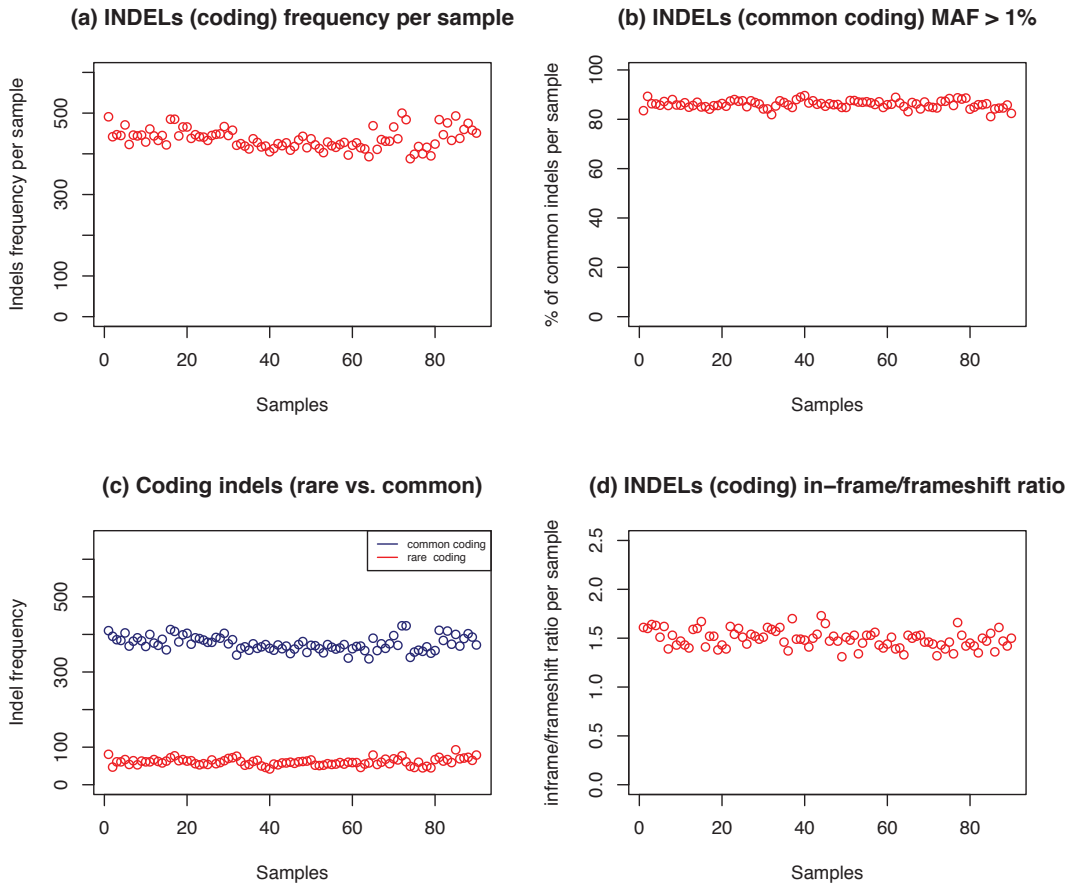


Figure 3-7 Quality control plots for insertion and deletion variants (indels)

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).

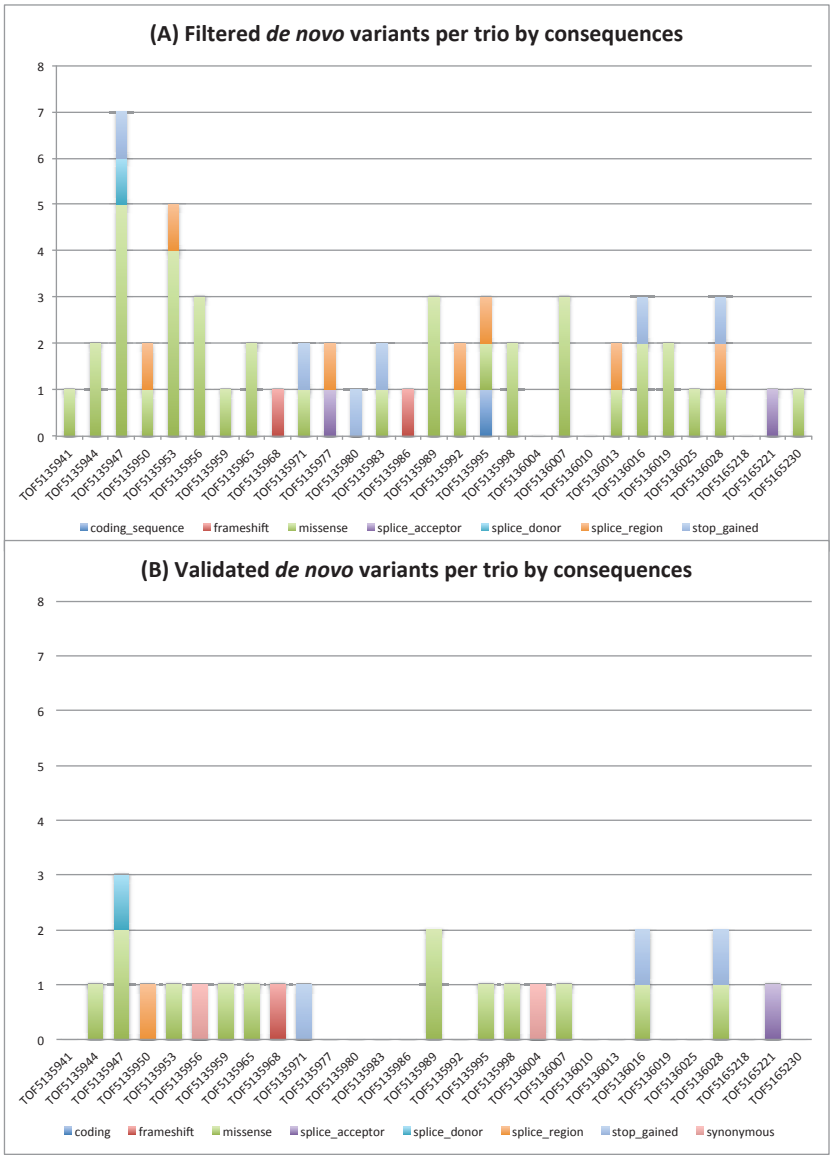


Figure 3-8 Filtered candidate *de novo* variants per trio by consequences. (B) Validated de novo variants by capillary sequencing.

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

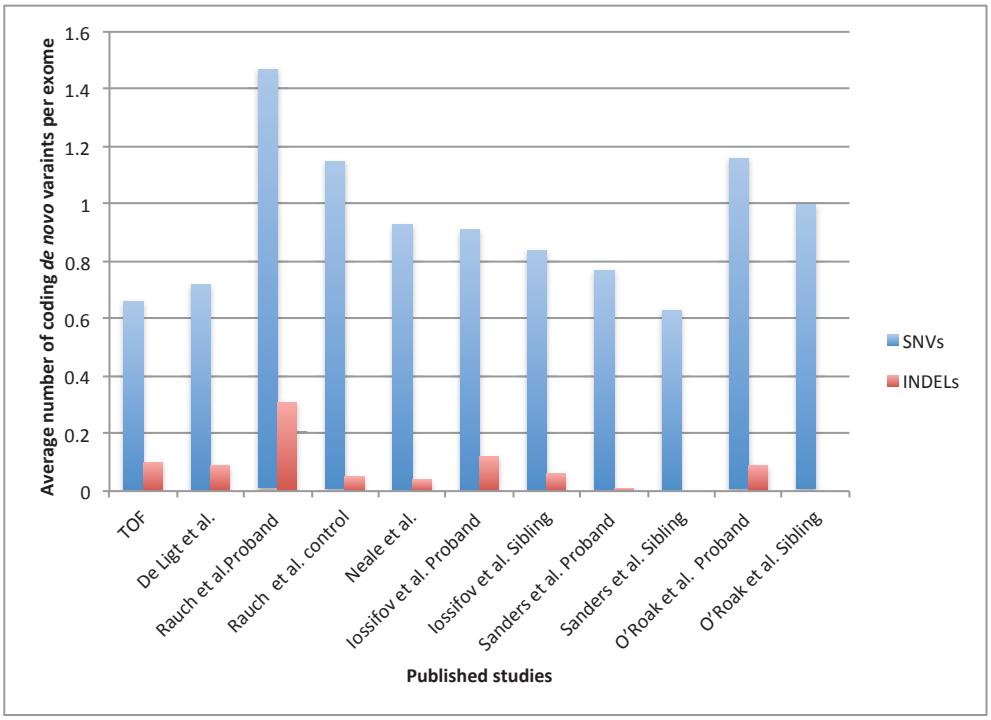


Figure 3-9 The average number of validated *de novo* in the primary ToF cohort is comparable to other published studies [190, 267-271]. (The literature survey is a courtesy of Dr. Matthew Hurles).

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).

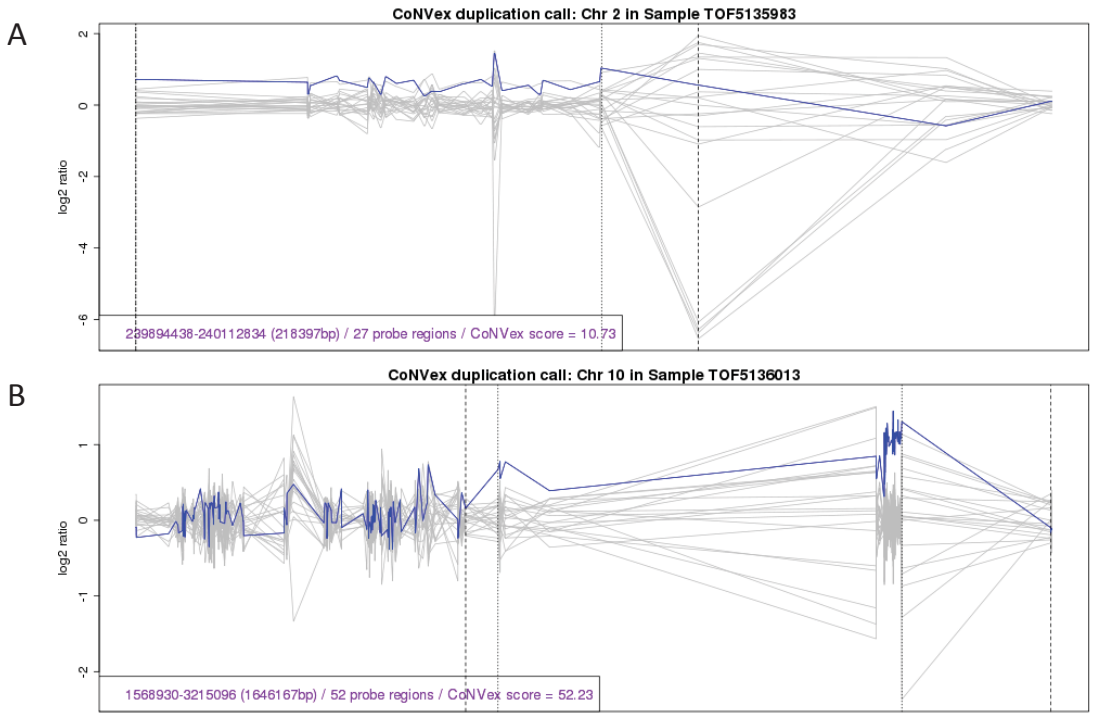


Figure 3-10: (A) A 218Kb duplication event on chromosome 2 spanning the *HDAC4* gene in patient (TOF5135983). The blue line is the log2 ratio in the patient while the grey lines represent the log2ratio scores for the same region in other samples in the cohort. (B) A 1.6 duplication event on chromosome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

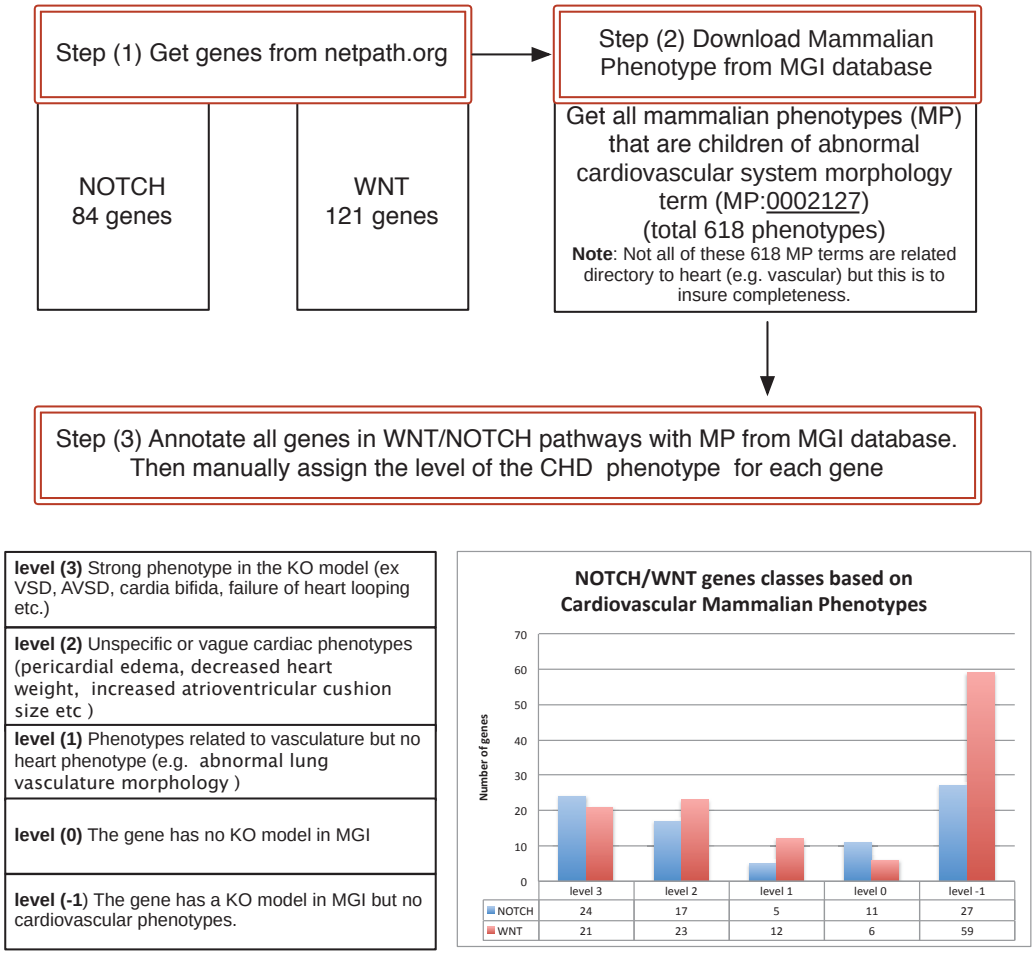


Figure 3-11 The workflow of gene selection from NOTCH/WNT pathway in the ToF replication study

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

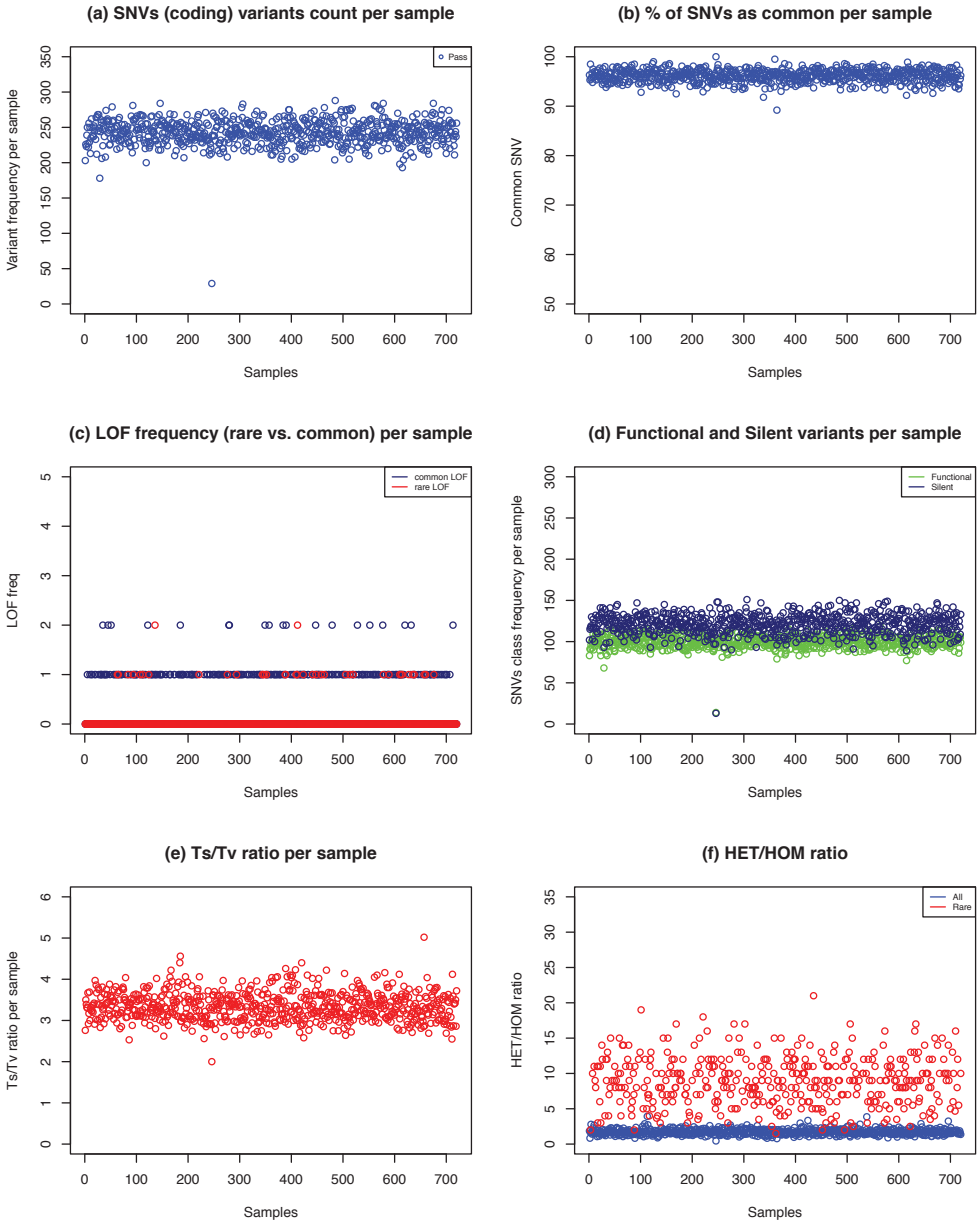


Figure 3-12 Quality control plots including global counts and various single nucleotide variants statistics in 209 trios from the ToF replication cohort (see main text for description)

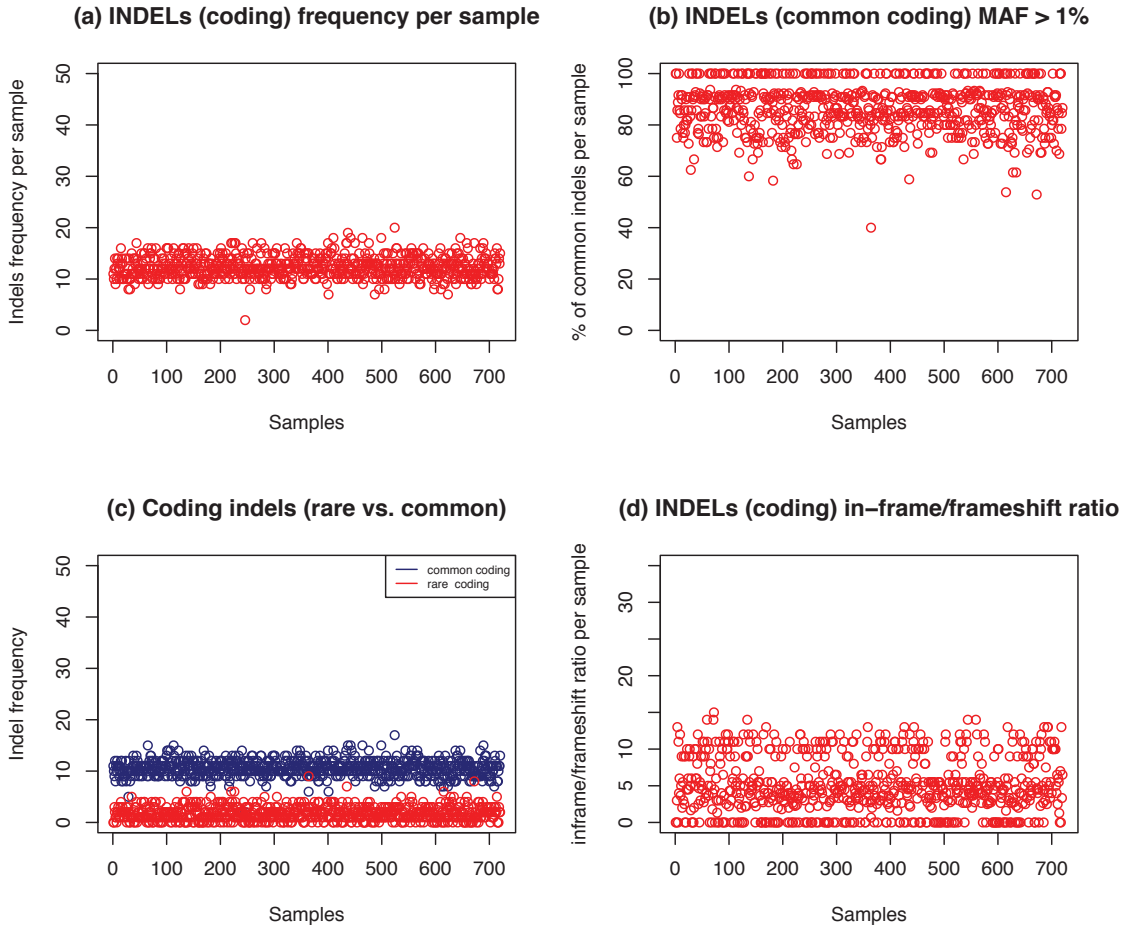


Figure 3-13 Quality control plots for insertion and deletion variants in 209 trios from the ToF replication cohort.

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.

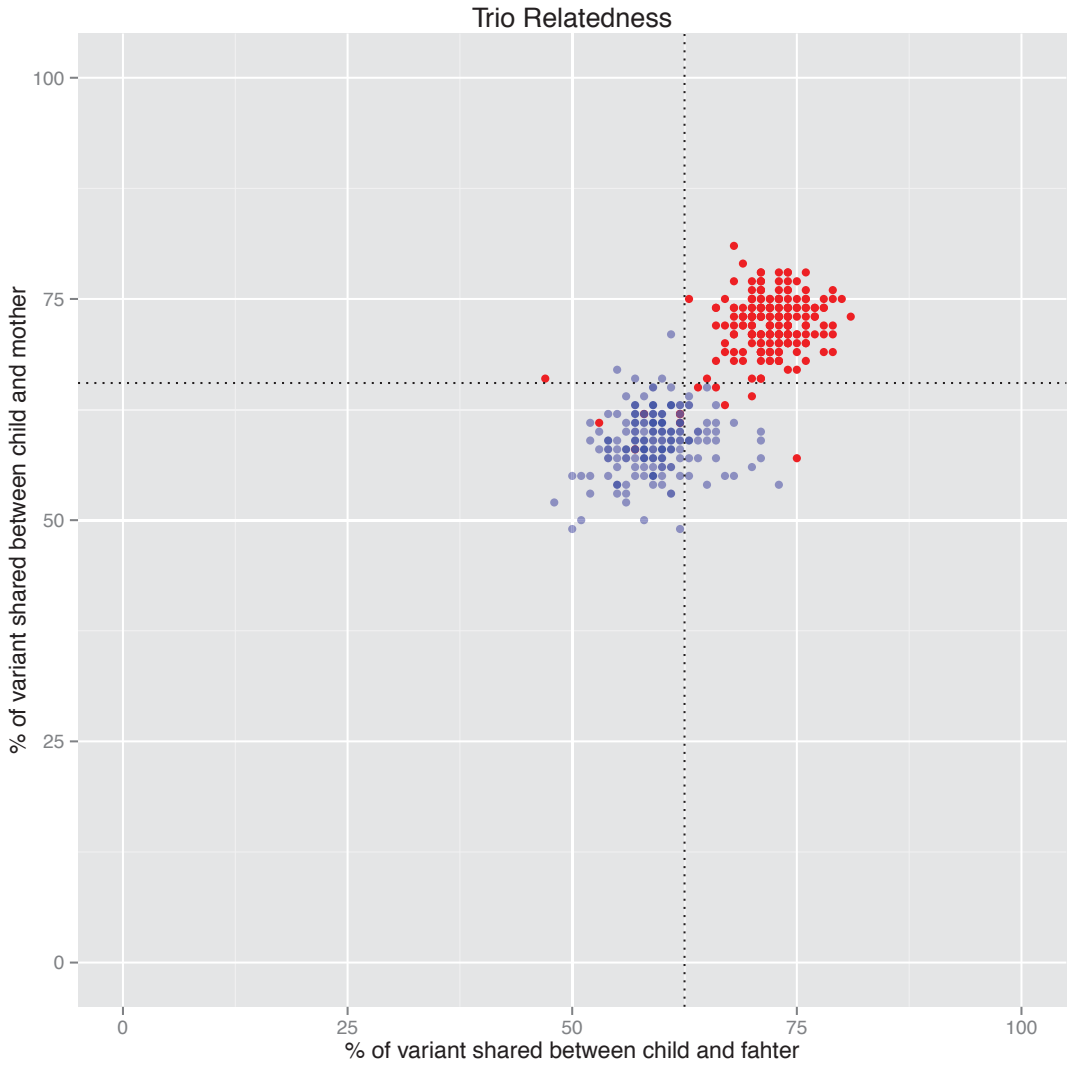


Figure 3-14 Percentage of shared variants between each child and his parents (red) and when children are assigned to random parent pairs (blue). Dashed black lines are used to separate the two groups and to flag six trios where children have shared < 62.5% of their variants with the father and/or < 65.5%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

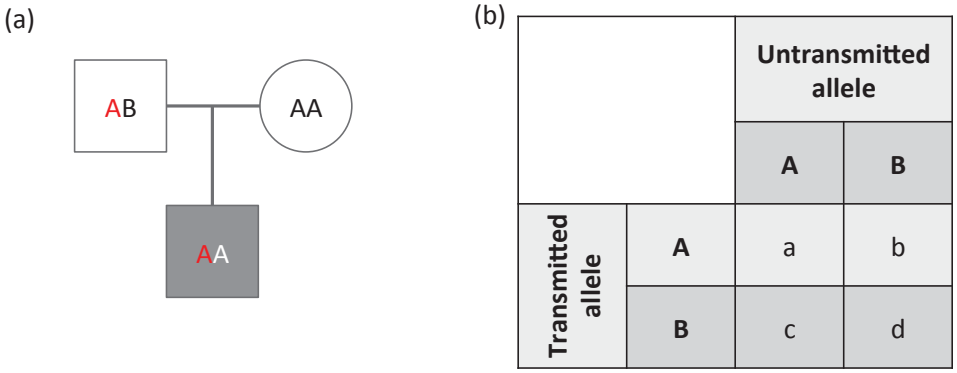
Sample ID	SC_RCTOF5364334		SC_RCTOF5394511			
Gender	Female		Female			
Chromosome	8	8	8	8	8	8
Position	144996830	145003613	144992962	144997315	144998052	144998495
Reference	C	C	G	T	T	G
Alternative	T	T	A	C	A	A
Gene	*PLEC*	*PLEC*	*PLEC*	*PLEC*	*PLEC*	*PLEC*
VEP predication	Missense	Missense	Missense	Missense	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

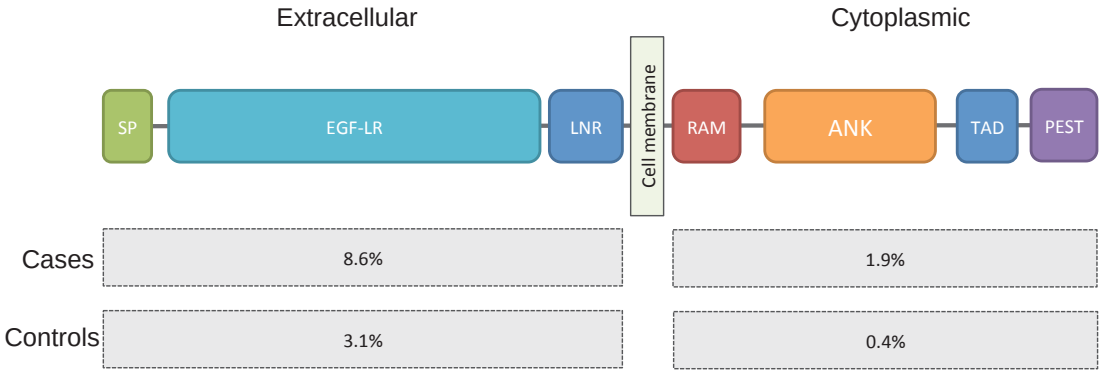


Figure 3-16 Mapping rare missense variants from cases (ToF replication cohort, n=209) and controls (DDD samples, n=1080) to the extracellular or cytoplasmic domains of NOTCH1. The majority of variants are in the extracellular domains where 8.6% of the cases has rare missense variants compared with 3.1% 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.

4 Combined genetic investigations of Atrioventricular Septal Defects (AVSD) in trios and index cases

Collaboration note

This chapter contains work performed in collaboration with many people, most notably Dr. Sebastian Gerety and Catherine Mercer. Sebastian performed the luciferase assays while Catharine mapped the exact locus of a de novo balanced translocation in a patient with coarctation of the aorta to NR2F2 (appendix B).

4.1 Introduction

Atrioventricular septal defects (AVSD), also known as 'common atrioventricular canal' or 'endocardial cushion defect', characterize a group of congenital structural defects in the atrioventricular septum of the developing heart. About half of AVSD cases are syndromic, mainly associated with Down syndrome where AVSD is thought to result from the overexpression of genes on chromosome 21 (see Genetic factors section below). However, the other half of AVSD cases is mainly isolated (patients without extracardiac phenotypes) and its genetic architecture remains largely unknown.

In this chapter, I describe how I used exome sequence data from non-syndromic AVSD cases from two different family-designs, trios and index cases, to discover genes enriched for rare, functional coding variants. Using this approach, I was also able to identify a novel gene, *NR2F2*, which causes AVSD and other CHD phenotypes in humans in a dosage-sensitive fashion similar to other key cardiac developmental genes such as *GATA4*, *NKX2.5* and *TBX1*.

4.1.1 Anatomical classification

The major hallmark of all AVSD is the common atrioventricular valve (AV) but AVSD subtypes vary with respect to the level at which shunting between the atria or ventricles takes place. The main two clinical AVSD subtypes are complete and partial (Table 4-1 and Figure 4-1). The complete subtype is characterized by a primum atrial septal defect (ASD) that is contiguous with a posterior (or inlet) ventricular septal defect (VSD), and a common AV valve. Typical partial AVSD is distinguished from complete AVSD by the absence of an inlet VSD. Another two types have been described: intermediate and transitional and both are considered subtypes of complete AVSD. In the intermediate subtype a bridging tongue of tissue divides the common AV valve into two distinct orifices. On the other hand, the transitional subtype has a small inlet VSD that is partially occluded by a dense tissue (chordal attachment to the septum) resulting in a defect that is similar to the physiology of a partial AV canal defect [434, 435].

Table 4-1 Anatomical classification of AVSDs

AVSD Types	Phenotype Components
Complete	**Balanced subtype** Complete failure of fusion between the superior and inferior endocardial cushions. Consists of *Primum ASD * Posterior (inlet) VSD * Common AV valve **Unbalanced subtype** In addition to balanced type defects in the balanced type. This type has hypoplasia in either the right or left ventricular.
Partial	Incomplete fusion of superior and inferior endocardial cushion and consists of: * Premium ASD * A single AV valve annulus with two separate valve orifices * Usually the anterior leaflet of the mitral valve is a cleft.
Intermediate	This is a rare form of AVSD that is similar to the complete AVSD * Large Premium ASD * Posterior (inlet) VSD But it also has a bridging tongue of tissue divides he common AVS valve into two distinct orifices. The intermediate and complete AVSD have the physiology and clinical features of an ASD and a VSD [434].
Transitional	Anatomically, it is subtype of the complete AVSD as it consists of: * Large premium ASD * Posterior (inlet) VSD * Cleft mitral valve But physiologically it is similar to the partial AVSD because of a dense chordal attachment to the VS that lead to small insignificant ventricular shunting and delineation of distinct left and right AV valve orifices. Both transitional and partial AVSD clinical picture of a large ASD.

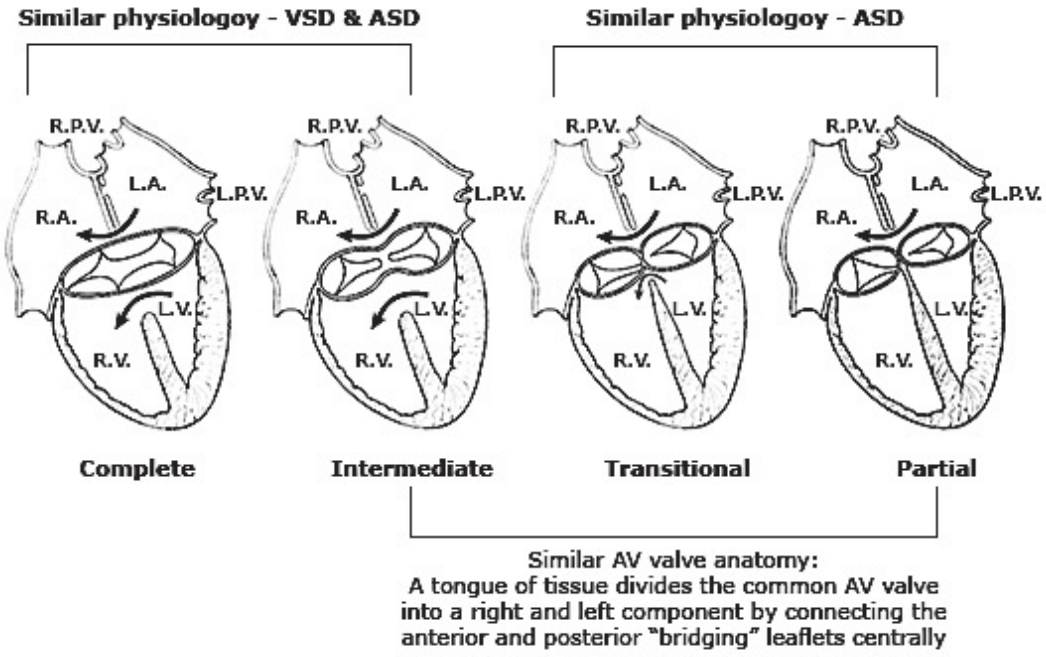


Figure 4-1 Anatomic and physiologic similarities between the different forms of atrioventricular septal defect (AVSD). Image adapted from [436].

The complete AVSD type is further subdivided using 'Rastelli classification' based on the atrioventricular valve morphology and the relative ventricular size [437]. The clinical severity varies depending on the size of the defect and whether it is associated with valvular defect and / or ventricular hypoplasia.

4.1.2 The prevalence of atrioventricular septal defects

AVSD represent 4-5% of all congenital heart defects (CHD) and its prevalence ranges from 0.3 to 0.4 per 1000 live births [438, 439] (Figure 4-2). However, AVSD prevalence is much higher in fetuses based on large fetal echocardiographic series where it was found to account for 18% of CHD cases [440]. The discrepancy in the prevalence may be attributed to the fact that many of the AVSD fetuses will not survive until birth either because they die prematurely or due to abortion. Postnatally, certain patient groups have a higher AVSD prevalence as in Down syndrome (44% of patients have CHD of which 39% are AVSDs) [311] and two-thirds of patients with heterotaxia exhibit one of the AVSD subtypes[441].

In a large population-based birth defects registry in Texas (USA), 1,636 cases of AVSD were reported between 2000-2009[442]. The most common AVSD subtype was complete AVSD (n= 1,335, 82%) [443]. More than half of the complete AVSD cases were syndromic (Table 4-2).

Table 4-2 The frequency of syndromic and non-syndromic complete AVSD reported between 2000-2009 in Texas birth registry **[443]**

Complete AVSD	n(%)
Syndromic	**772 (57.8)**
Trisomy 21	693 (51.9)
Trisomy 18	31 (2.3)
Trisomy 13	10 (0.7)
Other chromosome abnormalities	16 (1.2)
Other syndromes	33 (2.5)
Non-syndromic	**563 (42.2)**
Additional cardiac or non-cardiac malformation	516 (91.6)
Additional cardiac malformation only	223 (39.6)
Visceral heterotaxy	218 (38.7)

The recurrence risk (RR) of AVSD in first-degree relatives is 3-4% when one child is affected. While an affected father doesn't seem to increase the recurrence risk of AVSD, an affected mother, increases the RR up to 10% [15] (Figure 4-2-c). The male-to-female distribution of AV canal defect is approximately equal **[64, 444]** (Figure 4-2-d). Partial AVSD, however, shows a slight skew with more males affected than females (male-to-female ratio is 1.57) **[64]** but the small number of partial AVSD cases may explain this bias (n=18).

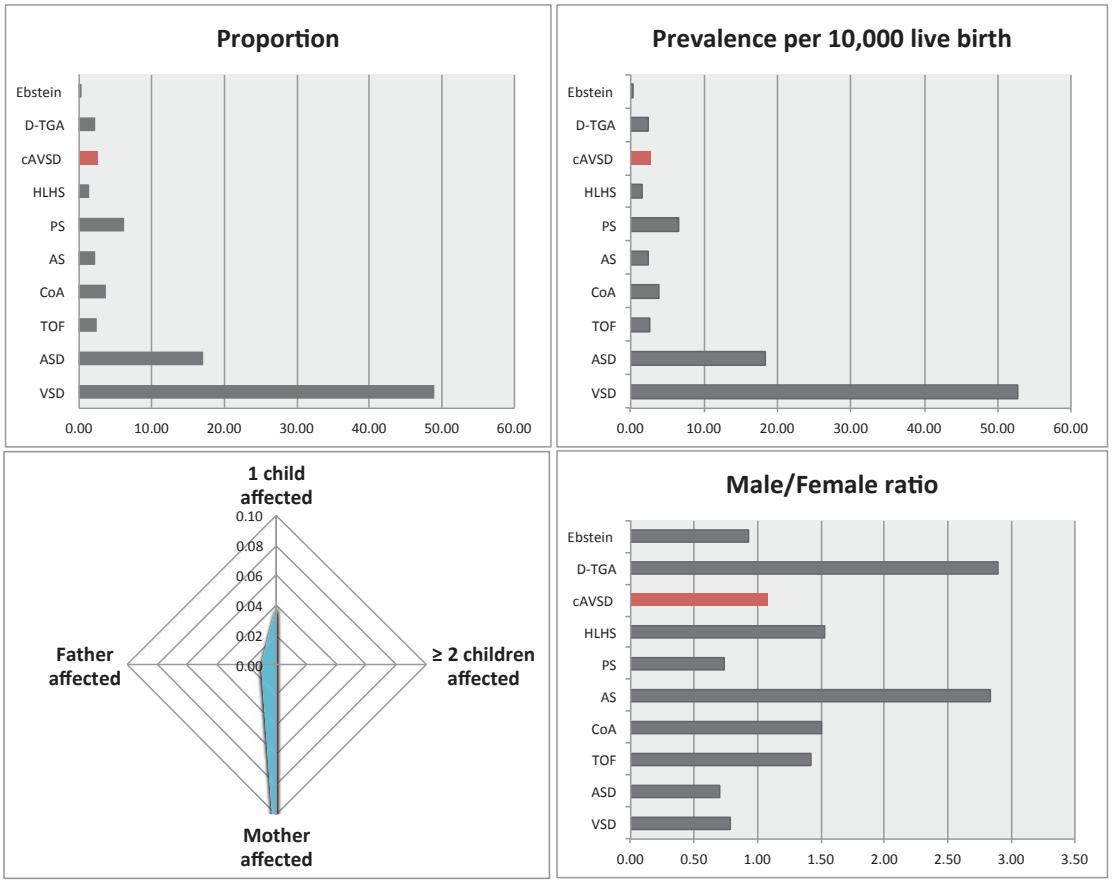


Figure 4-2 (a) proportion of different CHD, including complate atrioventricular septal defects (cAVSD) (red bar), in all cases registered in the PAN registery (n=7,245) during one year 2006-2007 (b) the prevealance of cAVSD cases in 10,000 live births from the PAN registry compared to other CHD cases (red bar). (c) Recurrence risk of cAVSD in first degree-realtives (d) cAVSD male-to-female ratio 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.

4.1.3 Clinical presentation

The clinical presentation of AVSD patients varies according to the size and extent of the defect and the presence of associated cardiac and/or extra-cardiac phenotypes. A newborn with complete AVSD may present with mild to moderate central cyanosis (bluish discoloration of the skin due to hypoxia) and develop congestive heart failure within a few months. The clinical examination may reveal a variable ejection systolic murmur, apical mid-diastolic murmur (in large left to right shunt), pansystolic murmur (with atrioventricular valve regurgitation). Additional tests are needed such as the electrocardiograph (ECG) to detect the presence of the superior frontal QRS axis, which is strongly

suggestive of AVSD, but chest radiograph and other advanced imaging approaches such as echocardiogram and magnetic resonance might be needed to confirm the clinical diagnosis [435].

Prolonged delay in surgical treatment may cause patients to develop Eisenmenger's syndrome that causes a permanent damage to the lung vascular circulation due to the long exposure to high blood pressure returning to the lung instead of the systemic blood circulation [445].

The prognosis of children with untreated complete AVSD is usually poor. Half of them die in the first year of life because of either heart failure or pneumonia. If they survive the first two years, an irreversible pulmonary vascular disease becomes increasingly common and affects virtually all patients [446]. The rate of 5-year survival is less than 4% in uncorrected complete AVSD patients [447]. However, long-term survival after surgical repair has been excellent and cumulative 20-year survival of 95% has been reported [448-450].

4.1.4 Embryological development of the endocardial cushions

The details of the development of the human heart have been described in chapter 1. This section summarizes the main events in the development of the atrioventricular cushion and related heart septation events.

At the ninth embryonic day (E9) of the developing heart in the mouse, the looped heart tube is segmented into four regions: the atrium, the atrioventricular canal (AVC), the ventricle and the outflow tract (OFT) (Figure 4-3). The heart tube is composed of an inner endocardial lining and an outer myocardial layer, which contain tissue swellings at the AVC lumen as well as in the proximal part of the OFT. These swellings are termed endocardial cushions and are formed by the accumulation of abundant extracellular matrix (cardiac jelly) inbetween the endocardium and myocardium.

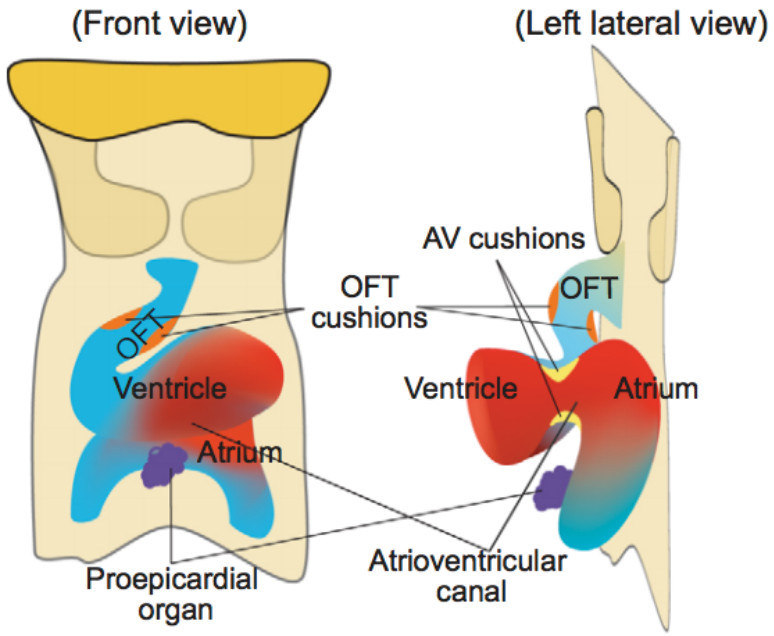


Figure 4-3 The formation of a mouse heart. Ventral and left lateral views at E9. The looped heart tube contains four anatomical segments: atrium, atrioventricular canal (AVC), ventricle, and outflow tract (OFT). Image adopted from [307].

For the AVC to develop into septal and valve tissues, its cushions require a population of mesenchyme cells. This population is derived through epithelial-to-mesenchymal transformation (EMT) from cells at the inner wall of the developing heart tube (endocardial cells). These endocardial cells differentiate into mesenchymal cells and migrate into the cardiac jelly to proliferate and form the AVC cushions [451]. In total, there are four mesenchymal tissues required for atrioventricular canal septation [307]: the superior and inferior atrioventricular endocardial cushions, the mesenchymal cap (MC), and the dorsal mesenchymal protrusion (DMP) [452, 453](Figure 4-4). The EMT process also is a key part of the mesenchymal cap (MC) growth from the lower part of the atrial septum [453]. The final mesenchymal set of cells required for AV canal septation in the dorsal mesenchymal protrusion (DMP) comes from the second heart field (SHF) which bulges into the atrial chamber as a mesenchymal protrusion [453, 454].

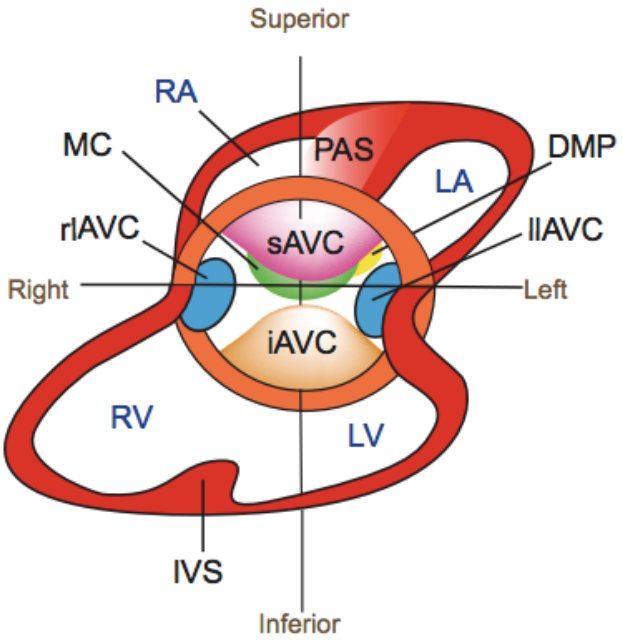


Figure 4-4 Superior and anterior oblique view of the AV cushion development. The AV canal will develop four cushions: the superior and inferior atrioventricular cushions (sAVC and iAVC) are the two major cushions in the central portion of the AVC and another two minor cushions, left and right lateral AV cushions (llAVC and rlAVC). The mesenchymal cap (MC) is a tissue that caps the leading edge of primary atrial septum (PAS) that grows from the atrial roof towards the AV canal. The dorsal mesenchymal protrusion (DMP) protrudes from the dorsal mesocardium into the atrial chamber. RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle; IVS, interventricular septum. (Adopted from [307])

These four mesenchymal tissues play a major rule in the septation of the AV canal in which any defect in the cellular migration and / or proliferation may cause atrial, ventricular or AV septal defects [307]. For example, the mitral and tricuspid orifices are separated when the mesenchyme of superior and inferior AV cushions fuses at the AV canal. A failure of the fusion between these cushions creates a common AV valve (AVSD). In a transverse section of the developing heart (Figure 4-5) the mesenchymal cap grows downward to reach and fuse with the AV canal anteriorly and creates part of the atrial septum. Similarly from below, an interventricular muscular septum emerges from within the ventricular chamber and grows superiorly to fuse with AV cushions, dividing the ventricular chamber into left and right ventricles [455, 456].

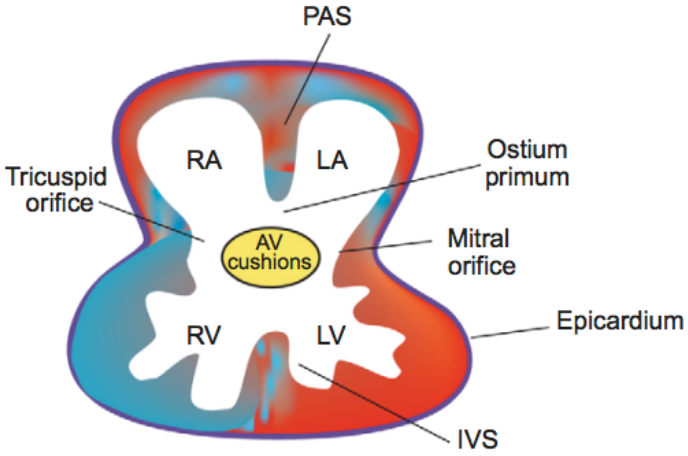


Figure 4-5 A transverse section at E11 in the developing mouse heart. At this stage, the heart is partially partitioned by the primitive atrial septum (PAS), interventricular septum (IVS) and atrioventricular cushions (AV cushions). The AVC is divided into tricuspid and mitral orifices, forming ventricular inlets that connect the respective atrium to the ventricle. The opening between the PAS and AVC is the ostium primum. RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle. (Adopted from **[307]**)

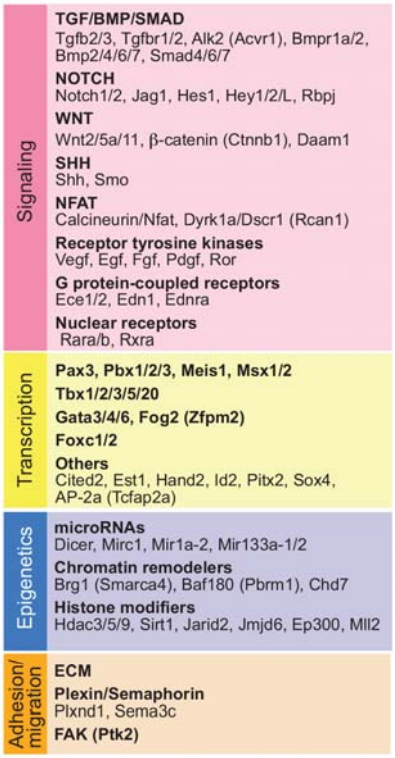


Figure 4-6 Genes and pathways essential for cardiac septation and valve development **[307]**

Studies of heart development in mouse models have linked 90-100 different genes in the regulation of heart septation and valve development (Figure 4-6). Broadly speaking, these genes can be arranged into four groups: signaling

pathways (e.g. NOTCH genes), transcription factors (e.g. GATA genes), epigenetic factors (e.g. microRNAs and histone modifiers) and adhesion or migration molecules. Many of these genes are discussed in chapter 1 and as part of different analyses in this thesis. Lin C. *et al.* have reviewed the role of these genes in much detail [307].

4.1.5 Causes of AVSD

4.1.5.1 Non-genetic factors

Many studies have addressed the involvement of environmental factors in the CHD (reviewed in chapter 1) but few have targeted non-genetic risk factors in AVSD specifically. The most detailed work in this regard was done in the Baltimore-Washington Infant Study [9, 297] where the authors detected many environmental risk factors for AVSDs such as maternal diabetes in non-syndromic AVSD infants (odds ratio=20.6). Maternal urinary tract infection was also found to increase the risk of AVSD, although mildly (odds ratio=2.29). Other AVSD risk factors are listed in (Table 4-3) along with their respective odds ratios and confidence intervals. Sonali Patel extensively reviewed the AVSD non-genetic risk factors extensively in her thesis [457].

It is important to note that these studies vary, and sometimes even contradict each other's conclusion. This can be attributed to the small sample sizes due to the rarity of AVSDs but also to the variation in the amount and length of exposure to these factors and how they were measured.

Table 4-3 Risk Factors and Exposures Associated With Atrioventricular Septal Defects

Condition	Risk Factor/Exposure	Odd ratio	95% Confidence intervals
Maternal Illness	Diabetes	22.8	7.4-70.5
	Urinary tract infections	2.29	1.11-4.73
Medications	Non-steroidal anti-inflammatory drugs (Ibuprofen)	2.49	1.42-4.34
	Antitussive medications	6.3	1.9-21.6
	Antibiotic medications	1.7	1.1-2.6
Non-therapeutic Drugs	Cigarette smoking (maternal)	2.50	1.21-5.19
	Cocaine	3.45	1.05-11.40
Occupational	Paint/Varnishes (maternal)	4.45	1.36-15.18

4.1.5.2 Genetic factors

4.1.5.2.1 Syndromic AVSDs

AVSDs can be part of syndromes caused by large chromosomal lesions, small microscopically visible events, or single point mutations. The Baltimore-Washington Infant Study (BWIS) identified 336 children with AVSD among 4,385 infants presenting under 1 year of age (7.7%) where 76% were syndromic [458], mainly Down syndrome (DS) [9]. In DS, 40-50% of the patients have CHD and the most common type is AVSD (of which 18% have a complete AVSD subtype) [311]. Having DS increases the risk of AVSD more than 2,000-fold [459]. The exact causes of CHD in DS are yet to be found, but many hypotheses have been suggested [460]. For example, overexpression of *DSCAM*, Down Syndrome Cell Adhesion Molecule, was suggested as the candidate of CHD in DS [461]. Similarly, *DSCR1* gene in the DS critical region is thought to disturb *VEGF-A*, an important regulator of endocardial cushions in the heart via the Calcineurin–NFAT pathway [104, 462].

Although having three copies of chromosome 21 genes increases the risk of AVSD and CHD in general, it is not sufficient to explain why half the DS patients have normal hearts. This has been suggested to be explained in part by the presence of rare deleterious coding variants in VEGF-A pathway genes (*COL6A1, COL6A2, CRELD1, FBLN2, FRZB*, and *GATA5*) in 20% of the DS cases (n=141) compared to 3% in healthy controls (n=141)[463]. This might indicate that the triple dosage effect of genes on chromosomes 21 may need a burden of rare coding variants to cause AVSD and other CHD but these findings have yet to be replicated by independent groups.

Other chromosomal lesions have been reported with AVSD. For example, distal deletion of chromosome 3p25-pter (3p− syndrome) causes low birth weight, mental retardation, telecanthus, ptosis, micrognathia, and AVSD in about third of the patients [464]. A consistent association was also described between 8p deletion (del8p) and AVSD [465, 466], which span a well-known CHD gene, *GATA4*. Additionally, there are a few reported cases of AVSD with partial 10q

monosomy, partial 13q monosomy, ring 22, 14q+, and 1p+3p- due to an unbalanced translocation [458].

Some Mendelian diseases caused predominantly by point mutations may present with AVSD. Two heterotaxy patients (OMIM 605376) with abdominal situs inverses and complete AVSD were found to have missense mutations in *NODAL*, a gene known to play a central role in early embryonic development, mesoderm and endoderm formation and left-right axis patterning [467]. Both recessive syndromes such as Ivemark syndrome (OMIM 208530), Ellis-van Creveld syndrome (OMIM 225500), Kaufman-McKusick syndrome (OMIM 236700) and dominant syndromes such as CHARGE syndrome (OMIM 214800) are also known to be associated with AVSD.

4.1.5.2.2 Non-syndromic AVSDs

Similar to other non-syndromic CHD phenotypes, the long-standing consensus on the genetic causes of isolated AVSD has focused on multifactorial inheritance, but this view has been challenged by the observation of several pedigrees with multiple affected individuals [468]. These findings suggested that a major genetic locus could account for the disorder in some families. Different loci have been linked to large families with isolated AVSD [469-474]. The common trend of these studies is autosomal dominant inheritance with incomplete penetrance and variable expression [475]. One of these loci associated with AVSDs is known as AVSD1 locus on chromosome 1p31-p21 (OMIM 606215), which was identified by use of a combination of DNA pooling and shared segment analysis in a high-density genome screen [476] but the exact causal gene has yet to be identified.

A second locus AVSD2 (OMIM 606217) was identified through analysis of chromosomal breakpoints in 3p- syndrome, which results from a deletion of 3p25-pter [464, 477, 478]. In this locus, *CRELD1* gene was proposed as the candidate gene for the AVSD2 locus on the basis of its mapping to chromosome 3p25 and its expression in the developing heart [479]. *CRELD1* encodes a cell

surface protein that likely functions as a cell adhesion molecule. A subsequent study by Robinson *et al.* showed rare heterozygous missense mutations in about 6% of isolated cases of AVSD in their cohort (two out of 35) [475] but further screening studies showed a lower rate of mutations in non-syndromic AVSD (ranged between 1.5 and 4% [480-482]. However, most of these studies lack functional experiments of compelling statistical enrichment to confirm whether these mutations are actually pathogenic or not.

The resequencing of known CHD candidate genes has also been used to look for rare coding mutations in isolated AVSD. Table 4-4 lists some of these genes along with the proportion of patients with rare coding mutations in every cohort. These studies, however, were able to explain only 2% of the isolated AVSDs on average. Another common feature shared between these studies was the lack of strong functional evidence for most variants. These factors, in addition to the incomplete penetrance and variable gene expressivity, make it hard to accept some of these genes as causes of isolated AVSD.

Table 4-4 Rare coding mutations detected in isolated AVSD candidate genes

Gene	Mutated patients / analyzed patients	%	Functional evidence	Reference
ALK2	2/190	1	Luciferase assay	Smith et al. [483]
ALK3	1/190	0.5	N/A	
ADAM19	1/190	0.5	N/A	
ERBB3	1/190	0.5	N/A	
EGFR	1/190	0.5	N/A	
UGDH	1/190	0.5	N/A	
FOXP1	1/190	0.5	N/A	
ECE2	1/190	0.5	N/A	
APC	1/190	0.5	N/A	
CRELD1	2/35	5.7	Western blot analysis (protein mobility)	Robinson et al. [475]
	1/49	2.0	N/A	Zatyka et al. [482]
GATA4	2/43	4.6	No mutation-specific assay (G4D mouse model)	Rajagopal et al. [484]
	1/190	0.5	N/A	Smith et al. [483]
	1/11	9.0	N/A	Zhang et al. [485]
GATA6	1/26	3.9	Luciferase assay	Maitra et al. [486]

4.2 Methods and Materials

Samples and inclusion criteria

Patients with atrioventricular septal defect (AVSD) without trisomy 21 or a *situs* anomaly, of Caucasian ancestry, with sufficient DNA available were included. Eligible patients underwent dysmorphology assessment and a review of medical records. Informed consent was obtained from parents/legal guardian.

Patients in the primary cohort were enrolled prospectively in different centers in UK, Europe and Canada. Our collaborators Seema Mital and Lisa D'Alessandro at the SickKids hospital in Toronto (Canada) selected about 60% (N=81) of the patients from an Ontario province-wide Biobank registry. Another 34 samples came from the Genetic Origins of Congenital Heart Disease (GO-CHD) collection by Shoumo Bhattacharya and Jamie Bentham (Oxford). A few additional samples (N=10) were collected at the Centre for Human Genetics, University Hospitals Leuven, Katholieke Universiteit Leuven (Belgium) by Koen Devriendt and Bernard Thienpont (Table 4-5).

The primary cohort includes 13 trios and 112 index cases of patients with different types of AVSD (Table 4-6). 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.

Table 4-5: The breakdown of AVSD subtypes in the discovery cohorts

AVSD TYPE	Cohorts			Total
	Leuven	Toronto	GO-CHD	
Complete	2	23	2	27
Intermediate	5	11	0	16
Partial	2	33	11	46
Unbalanced	1	11	0	12
Unknown	0	3	21	24
Total	10	81	34	125

Table 4-6: Family designs in the discovery cohorts

Family-design	Cohorts			Total
	Toronto	GO-CHD	Leuven	
Trio	3	0	10	13
Index	78	34	0	112
Total	81	34	10	125

Using the same inclusion criteria, the replication cohort included a total of 245 patients. Barbara Mulder collected 120 samples from the CONCOR-registry and DNA-bank, a joint registry of the Dutch Heart Foundation and the Interuniversity Cardiology Institute Netherlands (ICIN) of adults with congenital heart disease of Caucasian ancestry. Sabine Klaassen and her colleagues collected another 18 samples from the National Registry for Congenital Heart Defects, Berlin, Germany. The remaining samples were collected from GO-CHD and SickKids hospital (Table 4-7).

Table 4-7: The breakdown of AVSD subtypes in the replication cohorts (all are index cases)

AVSD TYPE	Cohorts					Total
	Berlin	CONCOR	Toronto	GO-CHD	Nottingham & Leicester	
Complete	6	14	2	80	2	104
Intermediate	7	0	1	0	0	8
Partial	5	105	1	11	4	126
Unbalanced	0	0	0	0	1	1
Unknown	0	1	1	0	4	6
Total	18	120	5	91	11	245

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 indel

were called using SAMtools and Dindel [158]. Variants were annotated for allele frequency using 1000 Genomes (June 2012 release), NHLBI-ESP (6503) project and UK10K cohorts. The Ensembl Variant Effect Predictor [170] was used to annotate the impact on annotated genes and GERP used for nucleotide conservation scores [165]. The variant calling and basic biological annotation of most samples were generated by the Genome Analysis Production Informatics (GAPI) pipeline (managed by Carol Scott *et al.*) except for 34 samples that were part of the UK10K RARE project, which went through UK10K pipeline (managed by Shane McCarthy *et al.*)[264]. Copy number variants were called using CoNVex pipeline by Parthiban Vijayarangakannan [372].

4.3 Results

4.3.1 Analysis overview

The main goal of my AVSD analyses was to identify genes with rare or novel-coding variants with a clear burden in cases compared with controls. This approach is based on a premise that part of CHD is caused by rare coding variants with large effect size (a monogenic model). However, this is hampered by the presence of many genes involved in heart development. Animal studies have identified hundreds of these genes and it is unlikely for any single gene to explain a large number of samples. On average, previous candidate resequencing studies had found rare coding variants in 2% of the patients (see Non-syndromic AVSDs section) assuming that we accept those variants as being genuinely pathogenic.

Figure 4-7 outlines the workflow and main analyses described in this chapter. The total number of isolated AVSD samples is 125; however, different pipelines were used to call variants in this cohort. Ninety-one samples went through the GAPI pipeline (the Genome Analysis Production Informatics, managed by Carol Scott *et al.*, described in chapter 2) and 34 samples went through the UK10K pipeline (managed by Shane McCarthy *et al.*).

Because the variant calling took place in two different calling pipelines, this led to some differences in the number of rare coding variants identified in each sample, which I described in chapter 2. Mainly, the number of loss of function variants in samples from UK10K is two times more than samples from GAPI pipeline. Additionally, the UK10K pipeline seems to under call rare homozygous coding variants as well as the coding INDELs in general. For these reasons, I decided to test two different sets of controls. The first set of control samples used for the rare missense burden analysis was obtained from the UK10K Neurological project (N=894) and all of these samples went through the UK10K pipeline. Later, I used a different set of controls chosen randomly from parental samples from the Deciphering Developmental Disorders (DDD) project (all from

GAPI pipeline) to see if changing the controls would improve the results burden of rare missense analysis.

To prioritize these genes, I used the *de novo* pipeline I implemented (described in chapter 2) to identify a list of genes with *de novo* coding variants and then intersect this list with genes from the burden analysis. The concept of narrowing down the search space for candidate genes using *de novo* analysis has been used successfully in Schizophrenia CNV studies (see for example [487]). Combining both *de novo* and burden analyses identified a single gene, *NR2F2*, which has one missense *de novo* variant in one trio and exhibit a burden of rare missense variants in another four cases (Fisher exact test P=0.00044). I increased the number of controls by including 4,300 samples from the NHLBI exome project (ESP) and was able to obtain a genome-wide statistically significant signal in *NR2F2* (Fisher exact test $P= 7.7 \times 10^{-7}$). I then attempted replication in a larger number of samples isolated AVSD cases (N=245) along with additional functional experiments to scrutinize the role that these variants may play *in vivo* and / or *in vitro*.

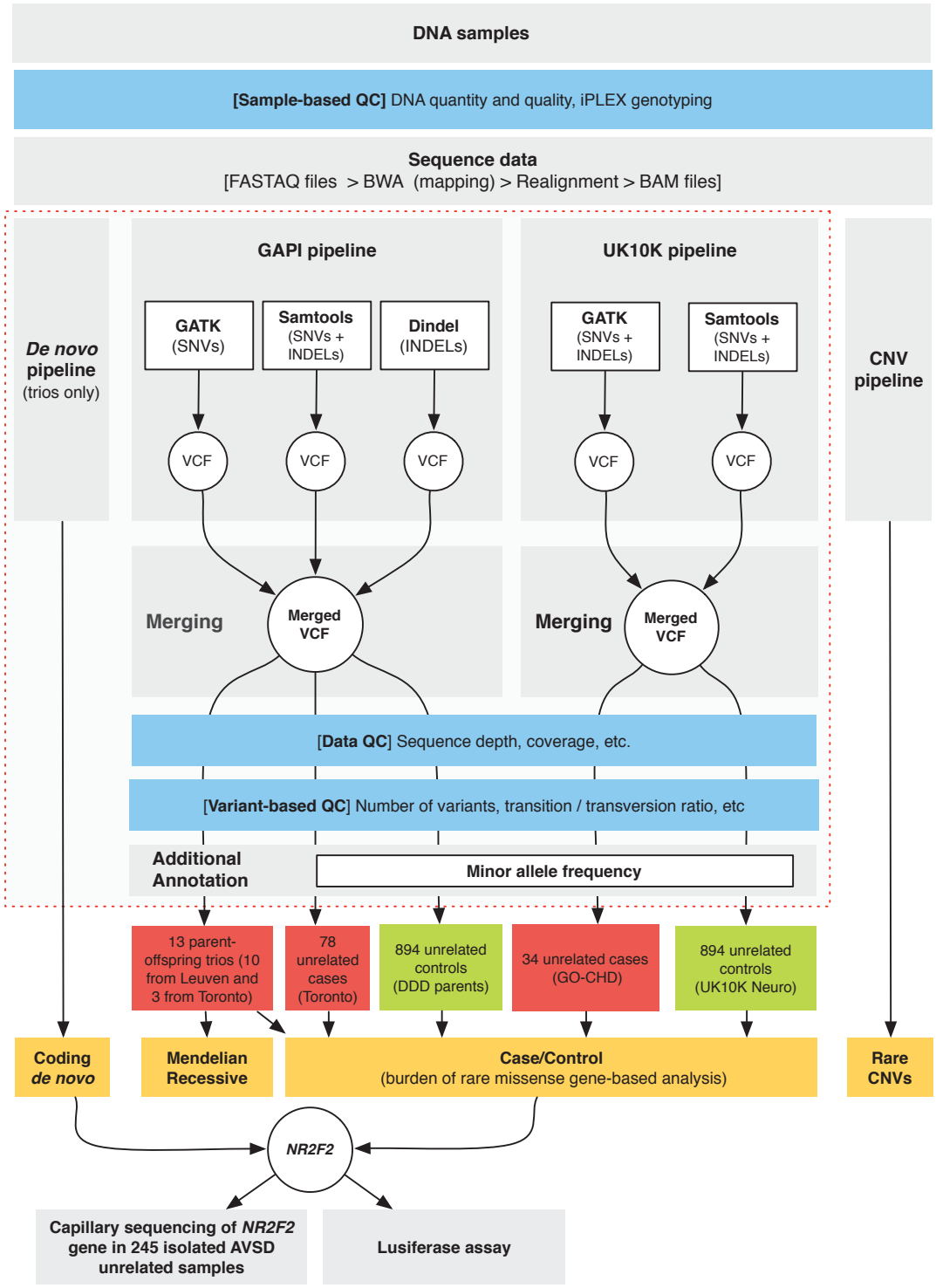


Figure 4-7 Overview of the workflow and analyses described in this chapter.
Red dashed box includes pipelines and tools that I described in chapter 2. GAPI: Genome Analysis Production Informatics, FEVA: Family-based Exome Variant Analysis, UK10K: UK10K variant calling pipeline. DDD: Deciphering Developmental Disorders (DDD) project, GO-CHD: Genetic Origins of Congenital Heart Disease sample collection (Oxford)

4.3.2 Quality control (QC)

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 sequence 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 contaminations issues. Only two samples were excluded from the AVSD cohort. The first sample had a degraded DNA (AVSD_1) while the second failed the gender matching test (AVSD_59). Both samples are part of the Toronto AVSD collection (Table 4-5).

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 and Shane McCarthy from the UK10K team have performed these tests to detect samples with too low sequence coverage. None of the cases failed any of these assessments. The average sequence data generated per exome is ~6 Gb with 65-fold mean depth and 85% of the exome covered by at least 10 reads.

DNA variant quality control

The third phase of quality control assesses the called variants in the Variant Call Format (VCF) files [161]. The aim of these tests is 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. Since AVSD samples belong to different cohorts, part of the samples went through the UK10K pipeline (mainly samples from the GO-CHD collection, n=34) while the rest went through GAPI pipeline (n=91 cases from Toronto and Leuven). Both pipelines used different variant callers (GAPI used GATK /Samtools to SNVs and Dindel/Samtools to call INDELs while UK10K used GATK/Samtools to call both SNVs and INDELs and did not include Dindel). Additionally, both pipelines used different number and variable thresholds to remove lower quality variants (full details described in chapter 2). These differences between GAPI and UK10K pipeline led to variability in the final number of coding variants (Table 4-8, Figure 4-8 and Figure 4-9). The most obvious three differences are the number of loss of function variants, the heterozygous/homozygous ratio for rare variants and the type and number of indels.

The UK10K pipeline called twice as many loss of function SNVs (LoF class includes stop gain and variant disturbing acceptor or donor splice sites) compared with the GAPI pipeline 188 and 93, respectively. However, I observed that most of the difference could be attributed to common LoF while both pipeline reported similar number of rare LoF (UK10K called 18 and GAPI called 14 LOF variants).

The second main difference I observed was the rare coding heterozygous/homozygous (het/hom) ratio (GAPI=7.4, UK10K=32.5). This big variation was not observed when I calculated the het/hom ratio for common coding variants (~1.5 in both pipelines). The main reason behind this variation is likely caused by UK10K under-calling rare homozygous SNVs. The rare heterozygous coding variants do not seem to be affected (the fraction of coding heterozygous variants that are rare in UK10K is 6.7% and 7.6% in GAPI). This

suggests the possibility of observing a false positive burden of rare homozygous SNVs when cases from GAPI are compared with controls from UK10K pipelines.

The third major difference in variants called by GAPI and UK10K is observed in indels. The GAPI pipeline calls 4.4x more coding INDELs than UK10K (462 in GAPI and 105 in UK10K). Additionally, the UK10K pipeline is enriched for rare indels in general (half of its coding indels are rare, < 1% MAF in 1000 genomes, compared to 18% in GAPI). Another difference is seen in the ratio of coding in-frame to coding frame-shift indels, which is used as an indicator of the calling quality of indels. As in-frame indels have a less severe impact, on average, on the protein structure than frame-shifting indels, we expect to see more in-frame due to weaker negative selection. Indels called by GAPI pipeline meet this expectation (coding in-frame/coding frameshift is 1.46) while UK10K show the opposite trend (ratio 0.44).

Using Dindel in the GAPI pipeline likely causes much of these differences in indel numbers. Dindel is a dedicated caller for indels that uses a probabilistic realignment model to account for base-calling errors, mapping errors, and for increased sequencing error indel rates in long homopolymer runs [158]. Dindel's superior performance comes at a price of high computation demands, which is why the UK10K informatics team has refrained from using it on large numbers of samples.

In summary, due to different workflows, variant callers and filters used by GAPI and UK10K pipelines, many important variations are observed in the number of coding variants. Indels in the UK10K pipeline exhibit strong differences that would certainly affect downstream analysis. SNVs on the other hand, are less affected than indels. Both pipelines show similar ratios of transition/transversion, heterozygous/homozygous, and rare/common variants. However, when I consider genotypes separately, the rare homozygous SNVs appear to be under-called in the UK10K pipeline.

Table 4-8 Quality control tests at different levels: sample-based, sequence data and variant-based levels. The most important variant calling differences between GAPI and UK10K pipeline are highlighted in red (rare heterozygous/homozygous ratio and in-frame/frameshift ratio for indels).

Stages	Goals	Tasks	Output	
DNA preparation	Amount and quality of DNA	Volume / concentration	All samples achieved the minimum requirement of whole exome sequencing	
		Genomic DNA integrity	1 sample excluded for degraded DNA (AVSD_1)	
	Quality assurance	Gender	1 sample excluded for gender mismatch with supplier sheet	
		Contamination	None of the cases show any contamination issues	
Stages	**Goals**	**Tasks**	**Average per sample (cases)**	
			GAPI (N=91)	**UK10K (N=34)**
Exome sequencing	Base-level stats	Raw output	~6 billion	~6 billion
		Average coverage per base	66	64
	Read-level stats	Raw read count	45 millions	44 millions
		Duplication fraction	6.8%	5.8%
Variant calling	Single nucleotide variants (SNVs)	Total number of coding SNVs	21,346	19,219
		Transition/Transversion ratio	2.98	3.12
		Heterozygous coding variant count (Het)	13,019	11,658
		Homozygous coding variant count (Hom)	8,326	7,561
		Het/hom ratio (all coding variants)	1.56	1.54
		% Of common coding SNVs (MAF > 1%)	94.9%	96%
		Common loss-of-function variants	79	170
		Common functional variants	9,569	8,829
		Common silent variants	10,185	9,361
		% Of rare coding SNVs (MAF< 1%)*	5.1%	4%
		Rare loss-of-function variants	14	18
		Rare functional variants	677	476
		Rare silent variants	357	257
		Heterozygous coding variant count (Het)	997	780
		Homozygous coding variant count (Hom)	134	24
		Het/hom ratio (rare coding variants)	7.44	32.5
	Insertion and deletion (indels)	Total number of coding indels count	462	105
		% Of common coding INDELs (MAF > 1%)	82%	49%
		Coding in-frame indels	274	33
		Coding frameshift indels	187	72
		Coding in-frame / frameshift ratio	1.46	0.45
		Rare coding indels	82	53

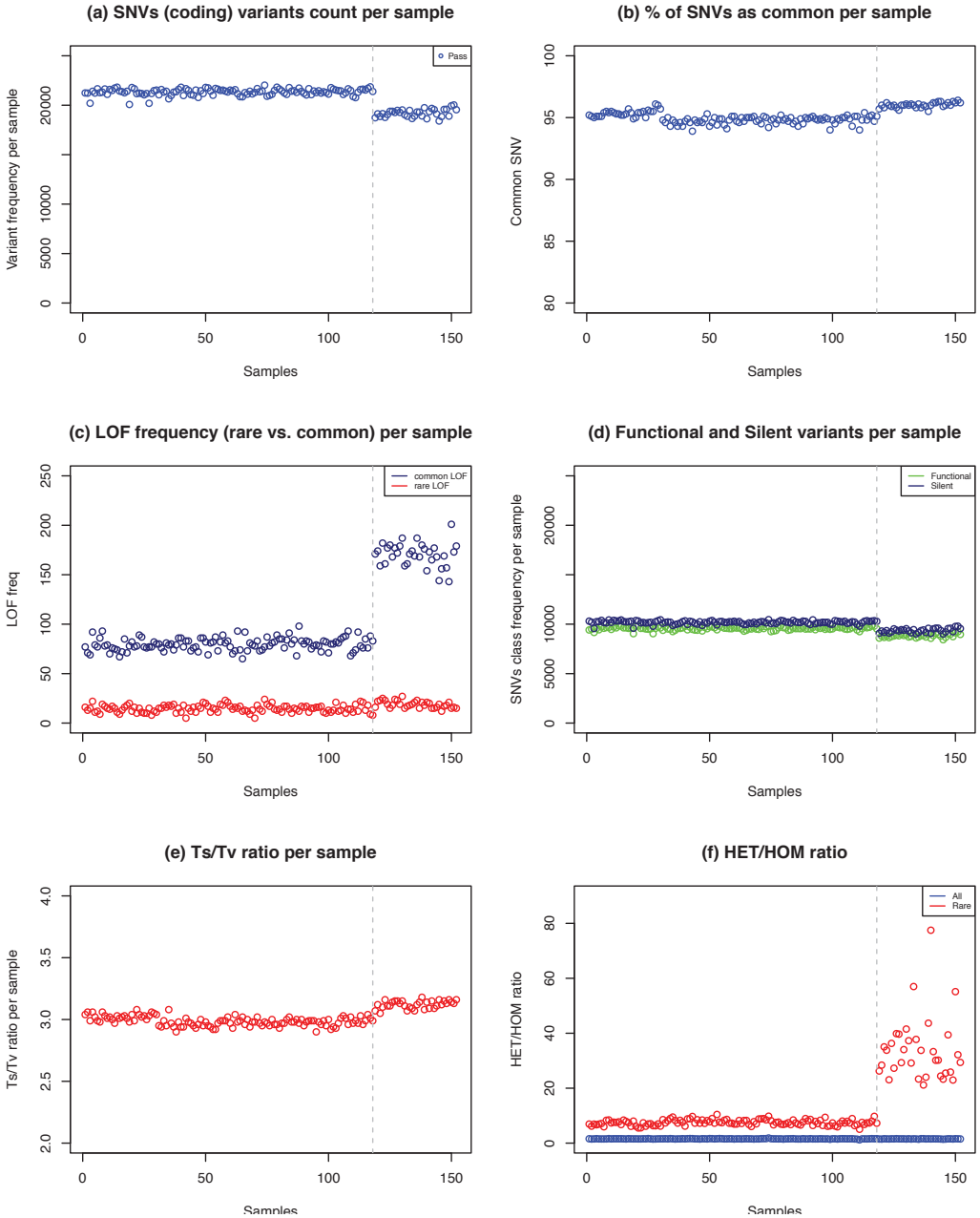


Figure 4-8 Quality control plots including global counts and various single nucleotide variants stats (see main text for description). Samples called by UK10K pipeline are plotted right to the dashed gray line. The remaining samples are called by GAPI pipeline.

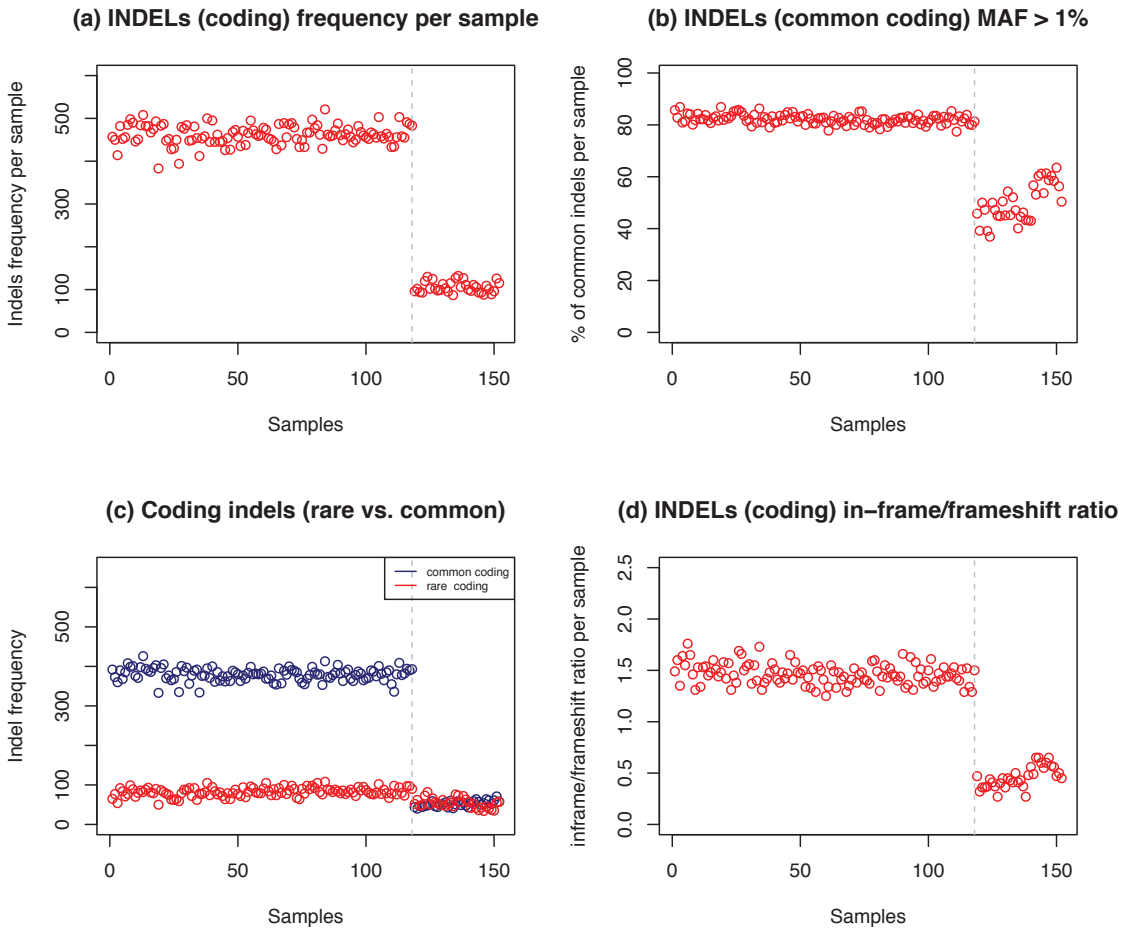


Figure 4-9 Quality control plots for insertion and deletion variants. Samples called by UK10K pipeline are plotted right to the dashed gray line. The remaining samples are called by GAPI pipeline.

4.3.3 Testing for burden of rare missense variants using controls from UK10K

The goal of this analysis was to look for the burden of rare missense variants in the cases (N=125 unrelated samples) compared with the controls. The controls I used were obtained from UK10K Neurological samples with the assumption that they do not exhibit any cardiac structural phenotypes. I selected 1,008 samples that are allowed to be used as controls. Before testing for the burden test, I needed to check for major confounding factors such as sample contamination, relatedness and population stratification that can easily cause biases in burden analysis and may generate false positive signals.

Exclusion of contaminated control samples

One of the quality control tests performed at the sample level (i.e. DNA) is genotyping 30-50 SNPs, which helps to detect gender mismatching and sample identification. However, sample contamination is harder to be detected at earlier stages especially if it is minimal or if the contamination takes place during library preparation and / or sequencing. The 1000 genomes project has used a program called "verifyBAMid" developed by Jun *et al.* at the University of Michigan to test for contamination issues using NGS data [488]. verifyBAMid checks whether the reads are contaminated as a mixture of two samples and generate a free-mix score. Shane McCarthy from the UK10K team generated free-mix scores and the het/hom ratio for all samples in the UK10K project including the UK10K neurological samples used as controls for this study (N=1,008). I plotted free-mix scores and the het/hom ratio for all samples (Figure 4-10), and used a threshold of 3% as suggested by verifyBAMid developers to detected possibly contaminated samples. This analysis identified 89 and I removed them from the downstream analysis.

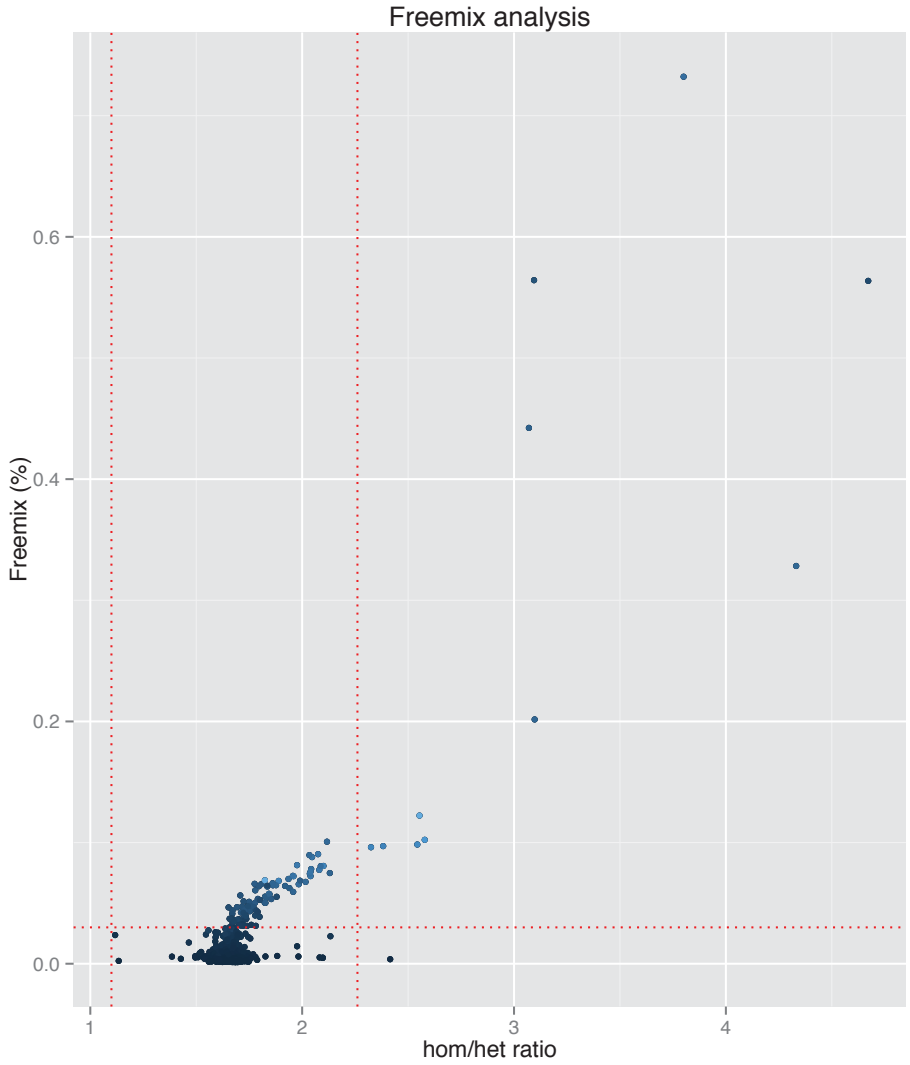


Figure 4-10: The heterozygous/homozygous ratio (X-axis) and free-mix fraction for 1,008 samples in UK10K neurological samples. The horizontal dashed red line is a cutoff 3% of free-mix suggested by the 'verifyBAMid' developers. Samples outside the two vertical dashed red lines at ±3 standard deviation of heterozygous/homozygous ratio were excluded. (Shane McCarthy provided the free-mix scores and het/hom ratios for the UK10K samples).

Population stratification

I used principle component analysis (PCA) to control for population stratification and make sure both cases and controls belong to the same population. All of the AVSD cases were recruited from Caucasian populations and I wanted to test if the control samples from the UK10K were also selected from the same population. I used 507 samples from four HapMap populations (African, Caucasian, Chinese and Japanese) as the reference populations for the PCA

analysis. First I selected extracted shared SNPs between HapMap samples and the samples from UK10K (n=69,415 SNPs) and removed non-autosomal SNPs, mutliallelic, rare SNPs with MAF < 5% and other steps (full workflow in Figure 4-11). These steps generated a high quality set of 10,492 SNPs to be used in the PCA analysis. This analysis showed that the majority of UK10K samples (n=919 controls and n=34 cases) overlapped well with European populations except for 25 control samples that I subsequently removed from any downstream analysis (Figure 4-12). Using the same workflow, I performed PCA analysis on the remaining samples from GAPI pipeline and all of the samples matched the HapMap Caucasian population (Figure 4-13).

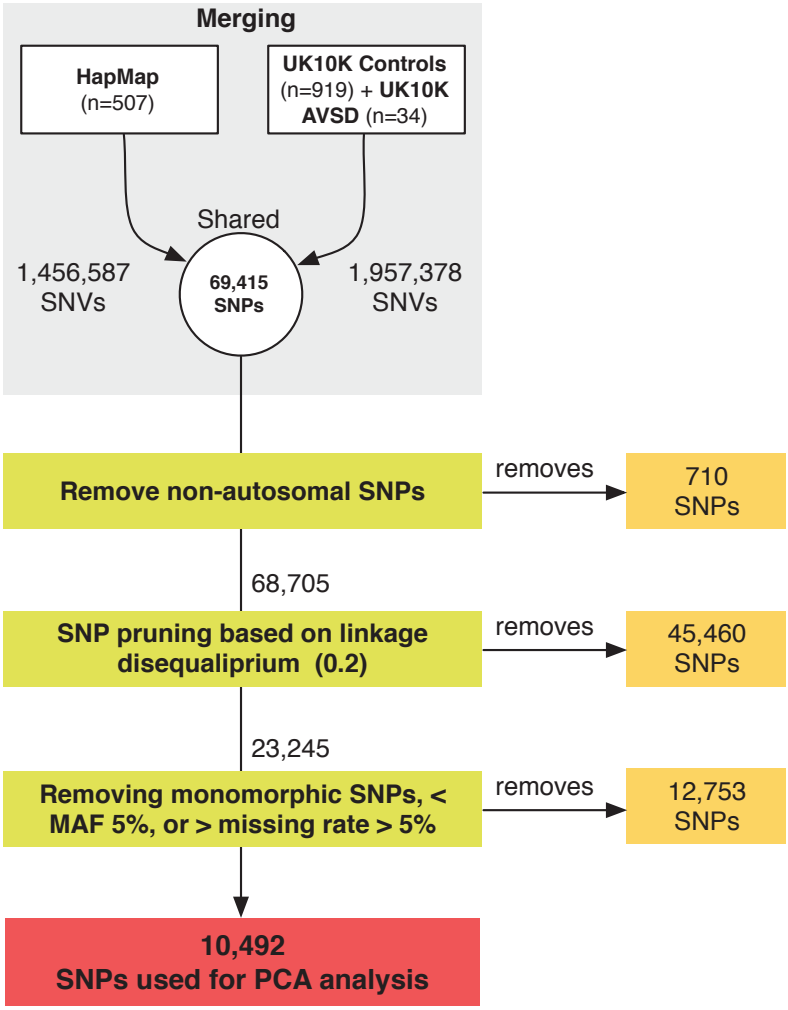


Figure 4-11 The workflow of SNPs selection for the principle component analysis (PCA). The reference SNPs are extracted from four HapMap populations (African, Caucasian, Chinese and Japanese) and found shared SNPs in 919 samples from UK10K control data. Similar workflow was performed for the cases as well.

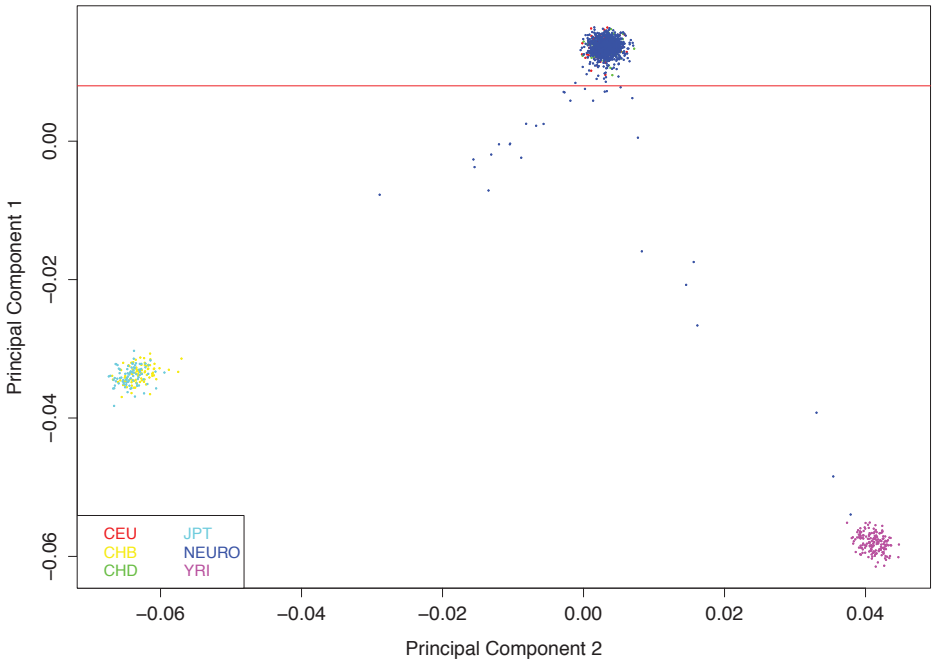


Figure 4-12 PCA analysis of 919 UK10K controls compared with main HapMap four populations. Control samples (UK10K) and AVSD cases from (GO-CHD) cohort. Twenty-five samples did not overlap with CEU population and therefore were excluded (blue points below solid horizontal red line)

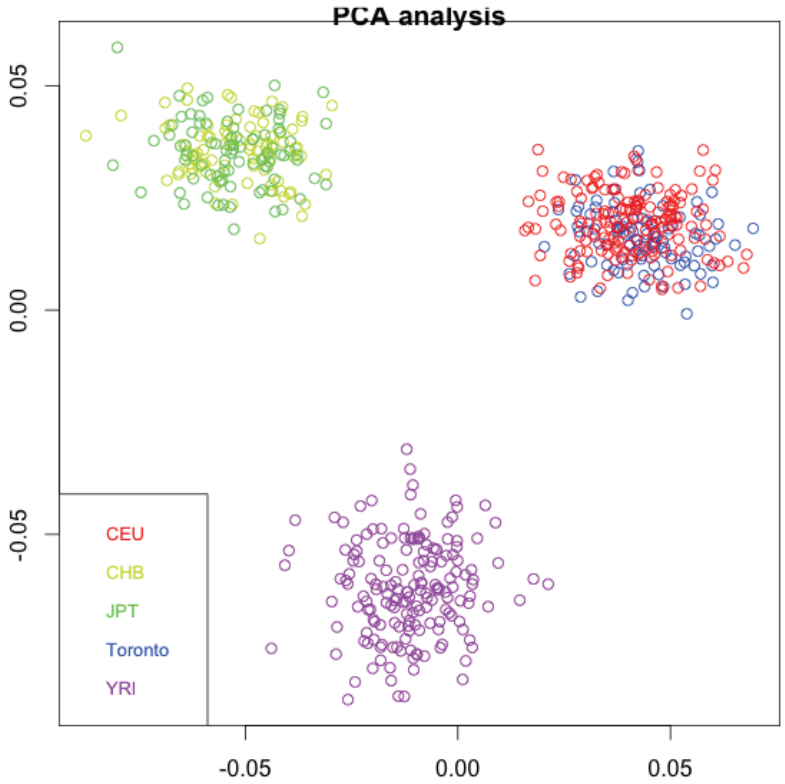


Figure 4-13 PCA analyses of the AVSD cases compared with the HapMap four main populations. The Toronto (AVSD) samples overlap completely with the Caucasian population. I have performed similar analysis for the remaining samples from Leuven (10 trios) and all of the samples overlapped with Caucasian population.

Collapsing rare variants per gene to increase the power of the test

To look for a gene-based burden of rare coding variants (except silent), I filtered out the common variants (MAF > 1% in the 1000 genomes or those that appear in > 1% of the in the cases and controls) and then grouped the variants by type (SNVs or INDELs) and variant consequences (loss-of-function or functional). The loss-of-functional class includes stop gain and variants disturbing donor or acceptor splice sites while the functional class includes the missense and stop lost variants. This was done separately for dominant (heterozygous) and recessive (homozygous or double heterozygous) variants. This arrangement generated four groups of candidate genes (Heterozygous-functional, Heterozygous-LoF, Homozygous-functional and Homozygous-LoF). Next, I created four 2 by 2 tables of the number of cases or controls that carry the variant in every group. Finally, I calculated the p-value using the Fisher's Exact test (right-tail only, since I am not looking for protective rare alleles). I decided not to include indels in this analysis given the big differences between GAPI and UK10K pipeline described above.

A common statistical approach used in genome-wide association studies to evaluate whether a statistical association test is generating unbiased p values is called the Quantile-Quantile (Q-Q) plot [489]. In QQ plots, the distribution of test statistics generated from the thousands of association tests performed (e.g. Chi square or Fisher exact test) is assessed for deviation from the null distribution (which is expected under the null hypothesis if no variant is associated with the trait).

Initially, I grouped AVSD cases from both GAPI (n=91) and UK10K pipelines (n=34) and compared them to controls from the UK10K pipeline (n=894). Figure 4-14 (plot A) shows the QQ plot for the burden tests of rare heterozygous functional variants in all genes. This showed an inflation of the observed p-values generated by the Fisher's exact test when compared with the null distribution on the x-axis. This is not unexpected given the known difference between the numbers of rare missense variants between the cases from GAPI

compared with controls from the UK10K pipeline (GAPI samples have 42% more rare missense variants per samples, see the variant-based quality control tests section above). To confirm this hypothesis, I decided to test the cases from GAPI and UK10K separately which, indeed, showed a worse inflation when using the GAPI samples alone (Figure 4-14, plot B) and improved when the cases and controls are both from the same pipeline (Figure 4-14, plot C and Figure 4-15).

Despite the slight improvement in the QQ plot when both cases/controls are from the same pipeline, the QQ plot is still showing signs of mild inflation (Figure 4-14, plot C). To see if the small number of cases (n=34) from UK10K caused this mild inflation, I increased the sample size by grouping all CHD samples I had from the UK10K pipeline (34 AVSD and 80 cases of mixed CHD subtypes, all unrelated) (Figure 4-14, plot D and Figure 4-15), which improved the QQ plot greatly.

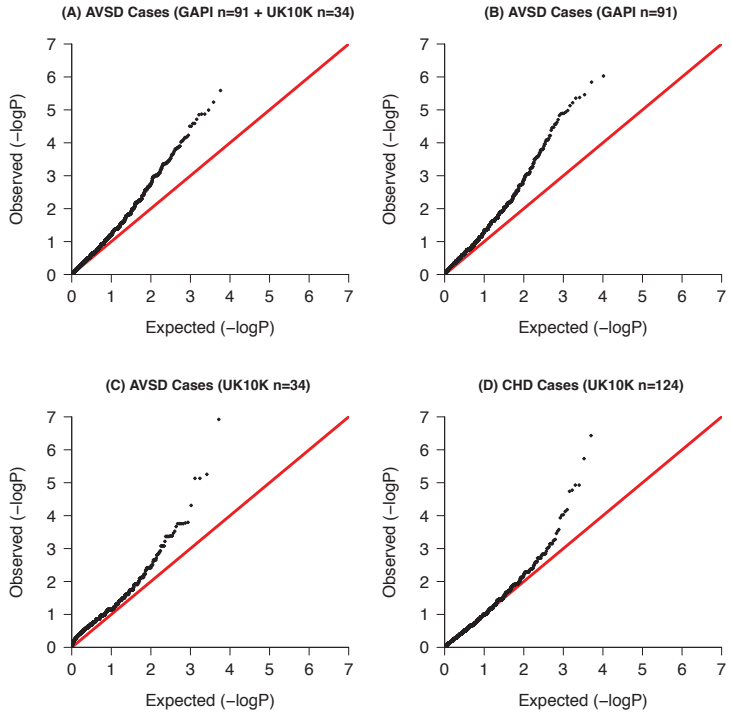


Figure 4-14 Quantile-Quantile (QQ) plots for the burden of rare heterozygous variant tests using four different sets of case samples. In all plots, the control samples are based on 894 samples from the UK10K neurological project. (A) QQ plot for 125 AVSD cases from both GAPI and UK10K shows marked inflation. (B) Same as plot A but includes cases from GAPI pipeline only which show worse inflation. (C) AVSD cases are limited to samples from UK10K only (n=34) which improves inflation since both cases and controls are from the same pipeline. (D) Represent the best QQ plot where, similar to plot C, both cases and controls are from the UK10K pipeline but I increased the number of cases by including all CHD samples from the UK10K pipeline (mixed phenotypes including the 34 AVSD cases).

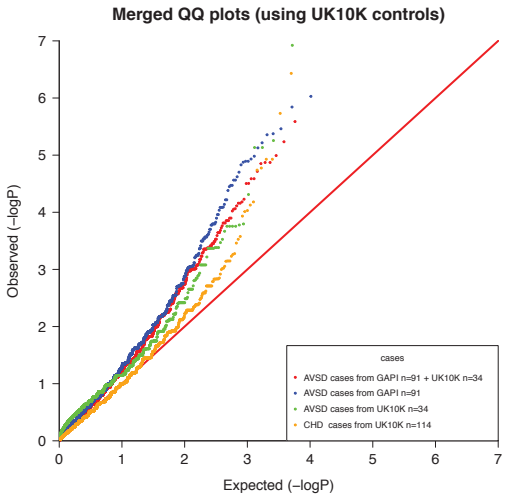


Figure 4-15 Combined QQ plots of four different sets described in Figure 4-14 to show the changes in QQ curves relative to each set. The most inflated set of cases is when I considered GAPI samples alone (blue) while the least inflated set is when I considered cases and controls from the same UK10K pipeline (orange).

Given the variability of QQ plots caused by combining the cases from different pipelines, I decided to use control data generated through the GAPI pipeline instead of the UK10K neurological controls to see if this would improve the QQ plots. I selected 894 parents at random from the Deciphering Developmental Disorders (DDD) project. Only one parent is selected from each trio to make sure I remove closely related parents. Using the same strategy described above, I grouped the AVSD cases into four sets: all AVSD from GAPI pipeline (n=91) and from UK10K (n=34) in one group, GAPI cases alone, UK10K cases alone and all AVSD with all other CHDs phenotypes we have sequenced so far as part of GAPI (n=263). The QQ plots (Figure 4-16 and Figure 4-17) show marked improvement over the QQ plots where I used controls from the UK10K pipeline. Besides changing the pipeline used to call control samples, increasing the number of cases from 91 AVSDs to 263 samples with different CHD subtypes also seems to improve the QQ curve (Figure 4-16, plot D).

Because most of the AVSD cases (n=91) went through GAPI pipeline, I decided to follow up the gene that shows a burden of rare missense compared to controls from the DDD (Figure 4-16, plot B). Table 4-9 lists the top 10 genes with significant p-values, however, after correcting for multiple testing only one gene shows a genome wide statistical significant p-value, *OR51E1*, which encodes for

an olfactory receptor and thus it is unlikely to be involved in the development of AVSD. Nonetheless, I used this list of genes to prioritize plausible candidate genes that I identified from subsequent analyses (e.g. *de novo* analysis).

Table 4-9 Top ten genes with a burden of rare missense variants in 91 AVSD cases from GAPI pipeline and 894 randomly selected parents from the DDD project used as controls from the same pipeline.

Genes	Samples with rare heterozygous missense variants					
	Cases AVSD (n=91)		Controls DDD (n=894)		Fisher Exact (right side)	Odds ratio
	Y	N	Y	N		
OR51E1	9	82	5	889	4.57E-07	19.51
PRPSAP1	6	85	1	893	3.46E-06	63.04
UCK1	8	83	7	887	1.48E-05	12.21
TMEM104	12	79	23	871	2.67E-05	5.75
LLGL2	13	78	28	866	3.12E-05	5.15
C6orf62	5	86	1	893	3.38E-05	51.92
TIE1	10	81	16	878	4.29E-05	6.77
PLEKHB2	8	83	10	884	7.94E-05	8.52
NR2F2	5	86	2	892	0.000109702	25.93
TOR2A	5	86	2	892	0.000109702	25.93

These results indicate that using samples from different pipelines is likely to confound the results of the burden of rare missense test and lead to either spurious association results. Nonetheless, despite the drawbacks of this combining of cases from two pipelines analysis, I coupled the results described here with the results from the *de novo* analysis to identify genes enriched in both analyses and then examined the burden signal in more detail using external control samples (e.g. data from NHLBI exome server) (see below section 4.3.5).

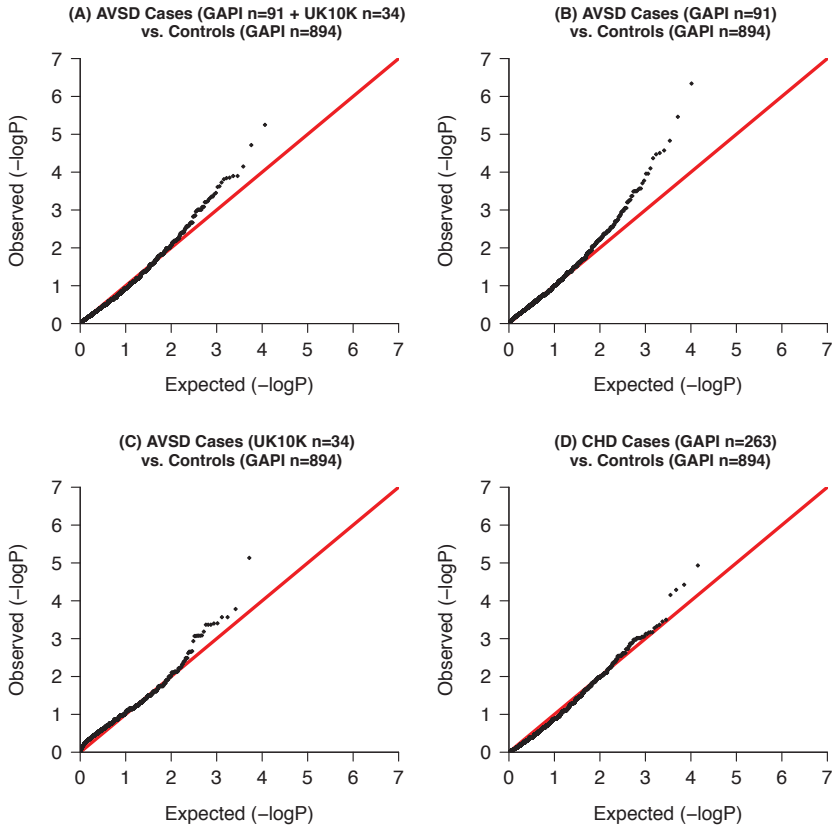


Figure 4-16 Quantile-Quantile (QQ) plots for the burden of rare heterozygous variant tests using four different sets of case samples. In all plots, the control samples are based on 894 samples from the Deciphering Developmental Disorders (DDD) project. (A) QQ plot for 125 AVSD cases from both GAPI and UK10K. (B) Same as plot A but include cases from GAPI pipeline only. (C) AVSD cases are limited to samples from UK10K only (n=34). (D) Both cases and controls are from the GAPI pipeline but I increased the number of cases by including all CHD samples from the GAPI pipeline (mixed phenotypes including the 91 AVSD cases).

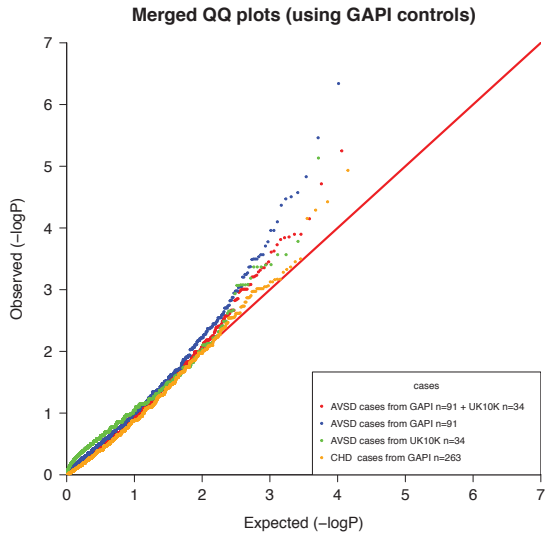


Figure 4-17 Combined QQ plots of four different sets described in Figure 4-16 to show the changes in QQ curves relative to each set.

4.3.4 *De novo* analysis

I used the DenovoGear (DNG) pipeline I developed previously (described in chapter 2) to detect candidate *de novo* mutations from the BAM files of 13 trios with AVSDs. On average, DNG was able to detect 180 potential *de novo* variants per trio. To minimize the false positive rate, I applied a few filters to exclude low quality, non-coding and / or common variants. These filters are (i) variant should not be in tandem repeat [490] or segmental duplication regions [491] from the UCSC tables[492], (ii) has minor allele frequency < 1% in the 1000 genomes, NHLBI-ESP (6503) and the UK10K cohort, (iii) fewer than 10% of the reads supporting the alternative allele in either parent (otherwise I considered it to be much more likely to be an inherited variant), (iv) variant should be called by an independent pipeline in the VCF file in the child but not the parents, and (v) the variant is predicted to be coding by VEP tool [170].

In addition to these five filters, DenovoGear software outputs a posterior probability score for each variant being a *de novo* (PP_DNM). This score can be used as an additional filter to reduce the number false positive rate. For example, removing variants with [<0.8] PP_DNM score increases the true positive proportion up to [80%] (personal communication with Aarno Palotie's team at WTSI). However, this strategy might be practical with a large number of trios (i.e. hundreds) but for small-scale project like AVSD trios, it is worth considering less stringent filters (I used the default PP_DNM > 0.001) to include the majority coding variants that pass the basic five filters above.

Figure 4-18-A shows the distribution of the plausible *de novo* candidates per trio after applying the basic filters (32 coding variants in total in 13 trios with an average of 2.4). I designed the primers for this validation and my colleague, Dr. Sarah Lindsay, performed laboratory work. Upon the analysis of the sequence trace files, I verified 40% of these *de novo* coding mutations (nine missense and four synonymous, Figure 4-18-B and Table 4-10) which lowers the average DNMs per trio to ~0.92. This average number of coding single nucleotide *de novo* variants corresponds well to other trio-based exome sequence projects such as

Tetralogy of Fallot trios (chapter 3) and other published studies (see *de novo* pipeline in chapter 2 for details) where the average of coding single nucleotide *de novo* variants of ranges (0.63-1.47). The remaining non-verified variants were either false positives (not present in any member of the trio) or inherited variants (present in both the child and one parent).

One trio in particular (CHDL5262758) carries four verified *de novo* mutations: two missense and two synonymous mutations. This is a rare event but still possible to observe. The frequency of *de novo* variants in large-scale projects tends to have a long tail of samples with more than one DNM (up to seven verified DNMs in DDD project, personal communication with Matthew Hurles).

The numbers of missense *de novo* variants are higher than the silent ones but the burden of *de novo* missense variants is not statistically significant. (exact binomial test, P= 0.77) compared with the expected proportion of *de novo* missenses by Kryukov *et al*. [357]. Only two genes with *de novo* missense variants show heart expression and / or a heart defect phenotype in mouse knockout mouse models (*NR2F2* and *ZMYND8*, Table 4-11).

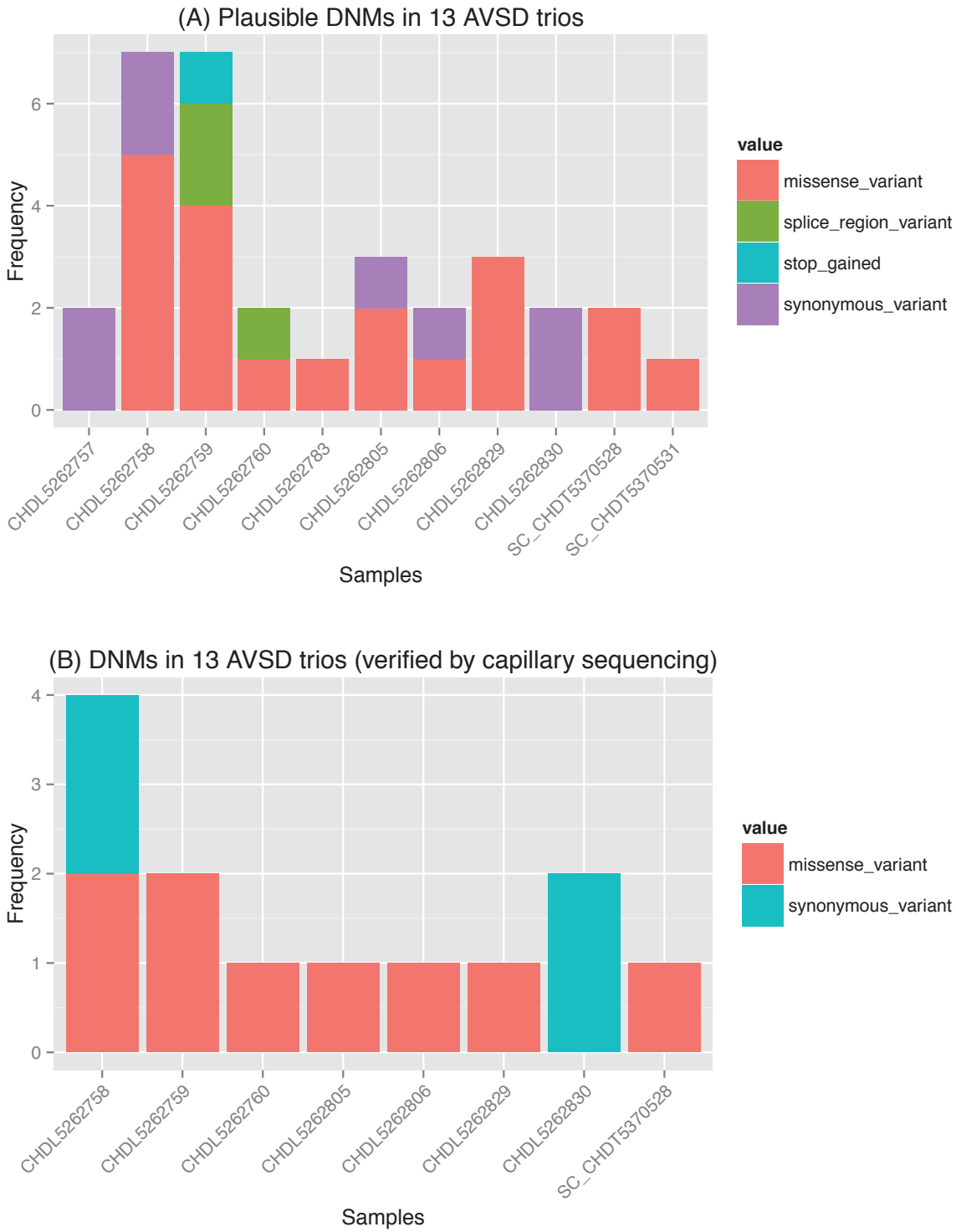


Figure 4-18 The distribution of the coding de novo mutation in 13 AVSD trios. (A) Plausible *de novo* mutations after applying five basic filters. (B) The distribution of verified *de novo* variants using capillary sequencing per trio. The variant predicted consequences on the protein are based on VEP program version 2.8. Only one potential loss-of-function variant appeared in *HDGFL1* but failed to validate in follow-up capillary sequencing.

Table 4-10: A List of verified coding DNMs in 13 AVSD trios.
REF: reference allele, ALT: alternative allele, PP_DNM: posterior probability of *de novo* variants.

Sample ID	CHR	Position	REF	ALT	PP_DNM	Gene	Predicted effect
CHDL5262758	1	225339733	G	A	1	*DNAH14*	Missense
	17	31323917	G	A	1	*SPACA3*	
CHDL5262759	20	61522324	A	C	0.386863	*DIDO1*	
	1	202129839	G	A	0.00998346	*PTPN7*	
CHDL5262760	2	80101311	A	T	1	*CTNNA2*	
CHDL5262805	9	84207971	T	C	0.00158238	*TLE1*	
CHDL5262806	2	190585499	T	C	1	*ANKAR*	
CHDL5262829	20	45927610	G	A	1	*ZMYND8*	
SC_CHDT5370528	15	96880628	C	A	1	*NR2F2*	
CHDL5262758	9	91994096	G	A	1	*SEMA4D*	Synonymous
	12	122396226	A	G	1	*WDR66*	
CHDL5262830	2	182394345	T	A	1	*ITGA4*	
	2	172650206	C	T	1	*SLC25A12*	

Table 4-11: The heart expression and phenotype in the knockout mouse models of the genes with verified functions *de novo* mutations

Candidate	Protein synopsis	Expression	knockout mouse model phenotype
SPACA3	Sperm surface membrane protein	No expression in the heart [493]	Not available
DNAH14	Ciliary dynein heavy chain 14	Undetected [494]	Not available
CTNNA2	Alpha-catenin-related protein	Mainly in the nervous system [495]	No, abnormalities of the brain includes a hypoplastic cerebellum [496]
DIDO1	Death-associated transcription factor 1	Undetected [494]	Anomalies in spleen, bone marrow, and peripheral blood [497]
PTPN7	Tyrosine-protein phosphatase non-receptor type 7	Undetected [494]	Mice homozygous for disruptions display a normal phenotype [498]
TLE1	Transducin-like enhancer protein 1	Expressed in adult heart, brain and kidney [499]	Not available
ZMYND8	Protein kinase C-binding protein 1	Expressed in multiple tissue including heart [500]	Not available
NR2F2	COUP transcription factor 2	Expressed in the mesodermal in most of developing internal organs [501]	Yes, atrioventricular septal defects in the conditional KO model [501]
ANKAR	Ankyrin and armadillo repeat-containing protein	Undetected [494]	Not available

4.3.5 Intersection between the results of the case/control and *de novo* analyses

To see if genes with *de novo* missense variants are enriched for rare missense variants, I intersected the results from both analyses (Table 4-12). Only one gene

in cases, *NR2F2* appears to be enriched for rare missense variants under the dominant model, when compared to controls with a p-value of ~ 1×10^{-4} (odds ratio of 18.6).

Table 4-12 The burden test rare missense variants burden in candidate genes obtained from the de novo analysis (i.e. each gene has at least one validated coding variants). Only one gene shows a significant burden, *NR2F2*.

Genes	Samples with rare Heterozygous missense variants					
	Cases		Controls		Fisher Exact (right side)	Odd ratio
	Y	N	Y	N		
NR2F2	5	86	2	892	0.00011	25.93
PTPN7	4	87	9	885	0.02545	4.52
ZMYND8	2	89	9	885	0.27006	2.21
TLE1	2	89	13	881	0.41049	1.52
DIDO1	6	85	44	850	0.31187	1.36
SPACA3	1	90	8	886	0.58362	1.23
CTNNA2	3	88	29	865	0.58093	1.02
SIK1	4	87	39	855	0.57453	1.01
DNAH14	5	86	64	830	0.78530	0.75
ANKAR	2	89	31	863	0.82697	0.63

To increase the power of the burden test, I included 4,300 European-American samples from the NHLBI-ESP project to the original control set (total n=5,194) [199]. However, the NHLBI-ESP project does not include sample-level genotypes. Instead, NHLBI-ESP provides alternative and reference allele counts for each variant in either African-American or European-American samples. I used this information to create a 2 by 2 table, similar to the sample-based burden test above, but instead of counting the number of samples, I conservatively assumed each alternative allele in the NHLBI-ESP set as an independent sample. Finally, I calculated the p-value of the burden test with Fisher's exact test.

Again, I found *NR2F2* to be the only gene with a significant enrichment of rare missense mutations but with more significant p value (P= 7.7×10^{-7}, odds ratio=54.1) (Table 4-13). This analysis detected two additional rare missense mutations in controls from NHLBI-ESP in addition to the original two missense

variants in the UK10K controls. Only one of the missense variants in patients (p.Ala412Ser) has previously been observed, in a single individual, in the 4,300 European-American exomes from the NHLBI-ESP project.

Table 4-13 The Burden test of rare missense variant in genes with confirmed *de novo* variants in AVSD cases compared to larger number of controls (NHLBI-ESP and UK10K Neurological control samples).

Gene	Cases (n=125)		Controls (n=5,194)		Fisher' exact *P*-value (two-tails)	Odds ratio
	With rare missense variants	Without rare missense variants	With rare missense variants	Without rare missense variants		
NR2F2	5	120	4	5,190	7.73E-07	54.063
ZMYND8	2	123	63	5,131	0.666	1.324
TLE1	2	123	64	5,130	0.668	1.303
PTPN7	4	121	137	5,057	0.574	1.220
DNAH14	11	114	302	4,892	0.174	1.563
CTNNA2	3	122	116	5,078	0.759	1.076
DIDO1	8	117	332	4,862	1.000	1.001
SPACA3	1	124	69	5,125	1.000	0.599
ANKAR	3	122	260	4,934	0.291	0.467

Since the exome sequence data in the NHLBI-ESP project was generated using smaller whole exome capturing kits (~17,000 genes compared to ~20,000 in my data), I examined the coverage and depth of sequencing of *NR2F2* gene in both cases and controls to investigate the possibility of variant under- or over-calling in cases or controls which can distort the results from the burden analysis. Figure 4-19 shows a comparable average depth per base pair across *NR2F2* gene in AVSD cases from GAPI and UK10K and the NHLBI-ESP control (UK10K=57x, GAPI=56x and NHLBI-ESP=67x). These analyses show that the coverage of *NR2F2* was very similar in the three pipelines and so the enrichment of rare functional variants in CHD is unlikely to be driven by technical biases.

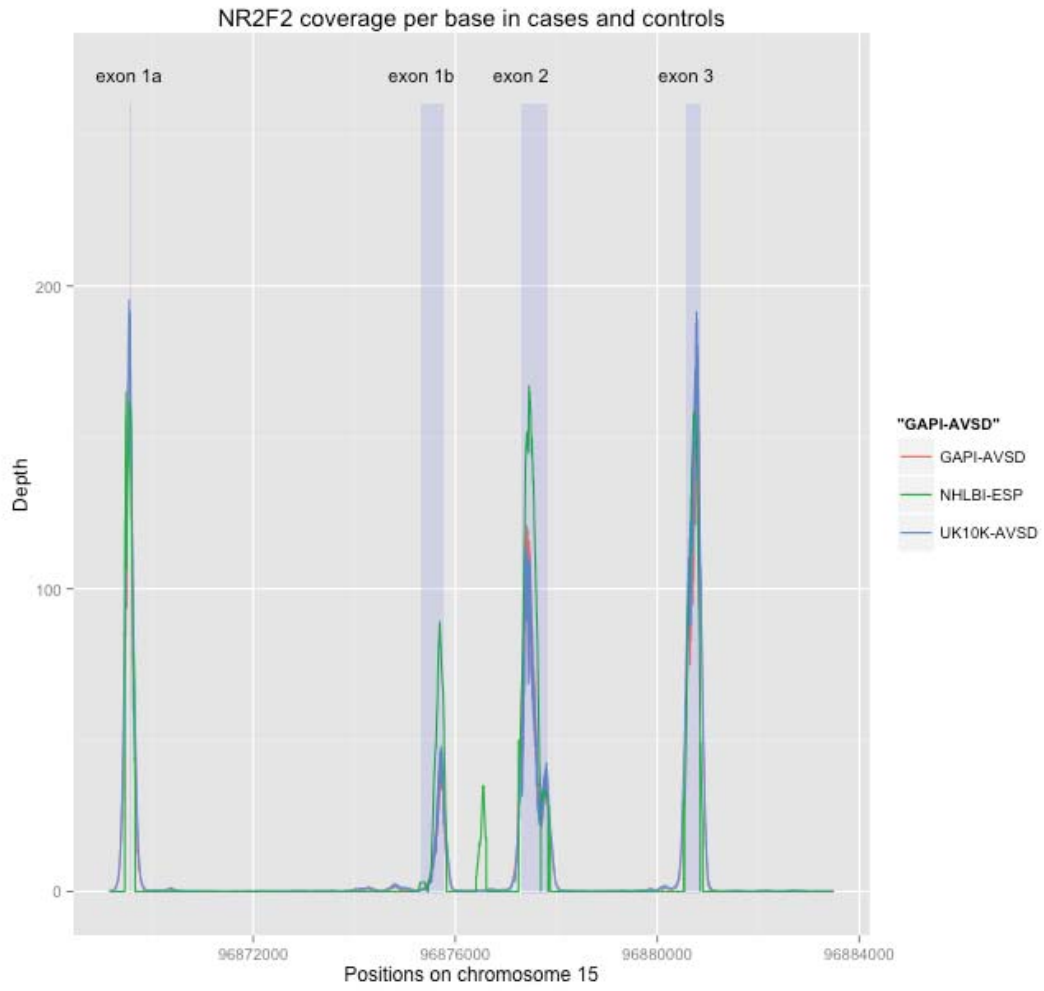


Figure 4-19 The average depth of *NR2F2* gene per base pair in the AVSD cases from GAPI and UK10K pipelines in addition to control samples from NHLBI-ESP project.

4.3.6 *NR2F2* mutations in the primary AVSD cohort

The AVSD analyses above identified only one gene, *NR2F2*, as a plausible AVSD candidate supported by evidence from two independent analyses: *de novo* analysis in AVSD trios and the burden test in the AVSD index cases. Five *NR2F2* rare missense variants were found in cases and four missense variants in controls (both UK10K and NHLBI-ESP sets) in this gene. One of the missense in cases arose *de novo* while the other four were in index cases. To determine the mode of transmission, our collaborators at the SickKids hospital Seema Mital and her team, contacted the families of the AVSD index cases. Three out of four families agreed to undergo a clinical examination and to provide DNA samples from the parents for validation by capillary sequencing. One variant,

p.Asp170Val also arose *de novo*, two of the other three missense variants observed in patients (p.Asn251Ile and p.Ala412Ser) were inherited from an apparently healthy parent (Figure 4-20-a and b), suggesting potential incomplete penetrance (capillary sequencing results are shown in Figure 4-25 b-f).

Moreover, the amino-acid changes observed in patients appear to be more disruptive than those observed in controls, as measured by the Grantham score, but with so few variants observed in controls, this trend is not statistically significant (Figure 4-20-c).

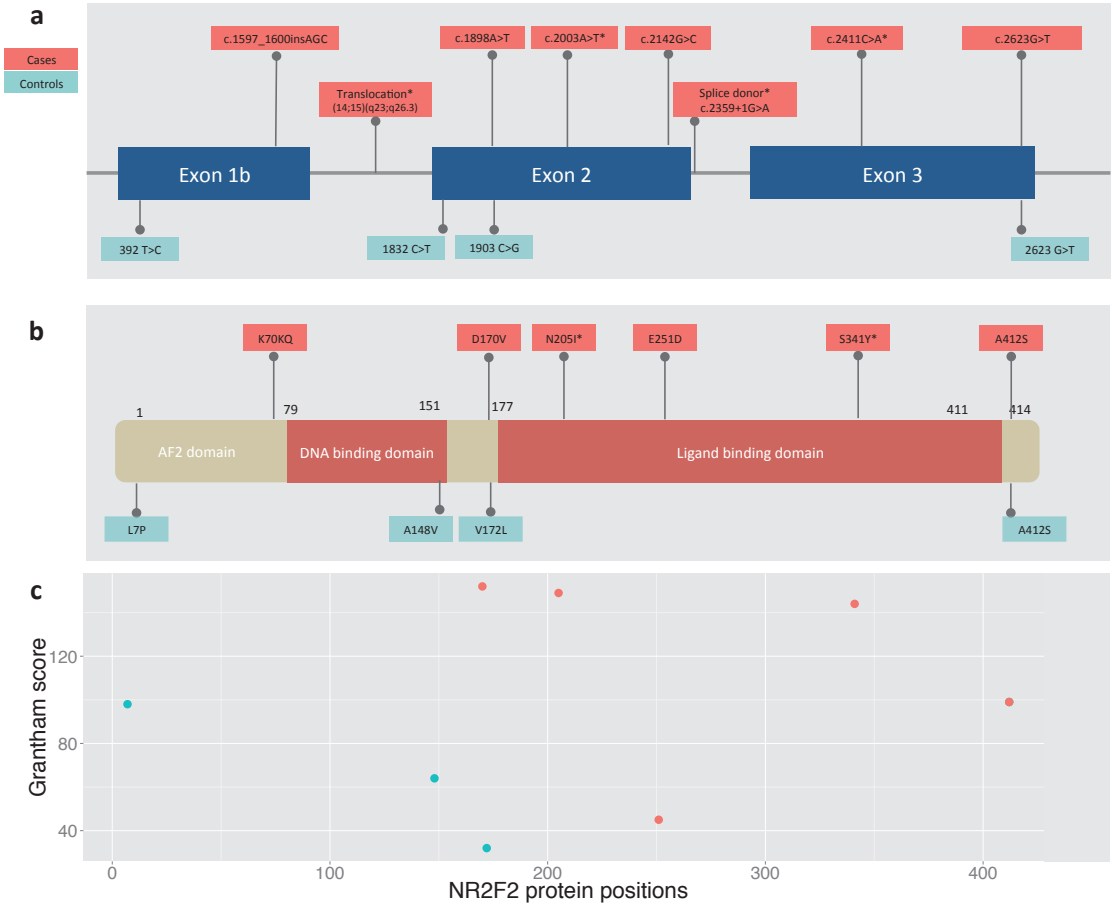


Figure 4-20 Structure of *NR2F2* gene and the encoded protein. (a) *NR2F2* gene has three coding exons and four transcripts. The transcript that generates the full-length protein (NM_021005) is shown here annotated with functional variants in cases (red) and controls (blue). (b) Similar to other nuclear receptors, NR2F2 has three main domains: a ligand-binding (LBD), DNA-binding (DBD) and an activation binding motif (AF2). Three mutations in cases are located in the ligand-binding domain (LDB). (c) The Grantham score for the missense mutations. *Denotes de novo variant

4.3.7 The effect of *NR2F2* mutations on the protein structure

The missense variants seen in patients are distributed throughout NR2F2, with three falling in the ligand-binding domain (p.Asn205Ile, p.Glu251Asp and p.Ser341Tyr). My colleague Jawahar Swaminathan was able to map two of these variants to a previously determined partial crystal structure for this domain [502] (Figure 4-21p.Asn205Ile is expected to perturb ligand binding whereas p.Ser341Tyr is predicted to destabilize the homodimerization domain).

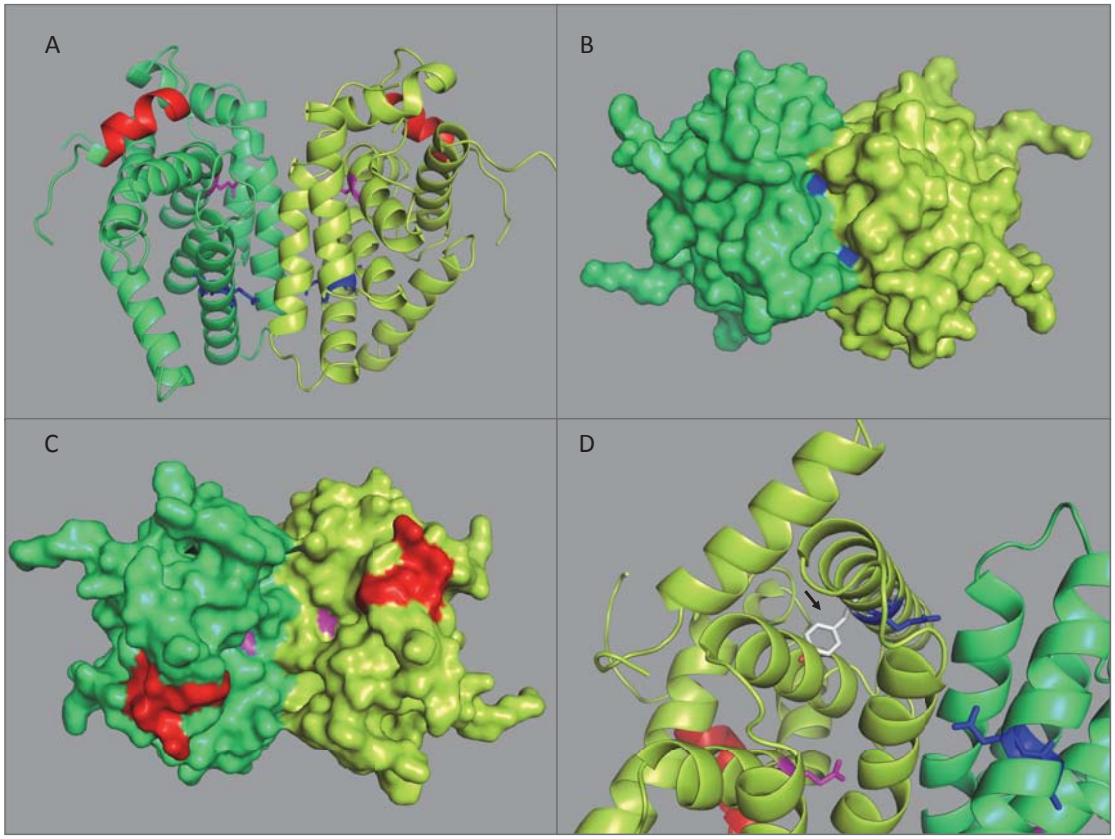


Figure 4-21 (A-C) Two missense variants mapped onto the partial crystal structure for the NR2F2 ligand-binding domain 10. p.Asn251Ile (purple) falls in the ligand-binding groove of the dimer, which in the repressed conformation is occupied by helix AF2 (red), and thus this variant is likely to perturb ligand-binding. p.Ser341Tyr (blue) is likely to destabilize helix A10 through steric hindrance and thus decrease the stability of NR2F2 homodimerization. (D) The *de novo* mutation (p.Ser341Tyr, blue color) effect on dimerization as it likely causes extreme steric hindrances that is likely to affect the critical dimer residue Q342 and helix A10 as a whole. This mutation will likely result in the movement of A10 and effect helices A7 and A8 as well.

4.3.8 NR2F2 exons and introns are very conserved

Nuclear receptor (NR) genes are generally conserved but the COUP-TF, NR2F2's gene family, is the most conserved NR family. For example, the ligand-binding domain DNA sequence of *NR2F2* or *NR2F1* is 99.6% similar between vertebrates and > 90% similar compared to *Svp* gene, the COUP-TFs homologue in the arthropod *D. melanogaster* [503]. Figure 4-22 shows high GERP [165] scores, not only in the exons but also within *NR2F2* intronic regions and extends to the flanking regions. The average GERP score per gene length ranks *NR2F2* in the top 10% of all genes (Figure 4-23). This high level of conservation of *NR2F2* domains between different species indicates very important biological functions and may explain why we observe very few missense variants in *NR2F2* across thousands of controls.

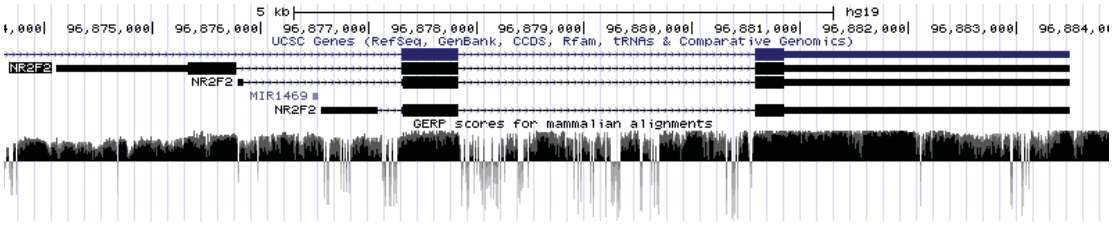


Figure 4-22 GERP scores per single base across NR2F2 (UCSC genome browser) showing high conserved scores in exons, introns and the flanking regions.

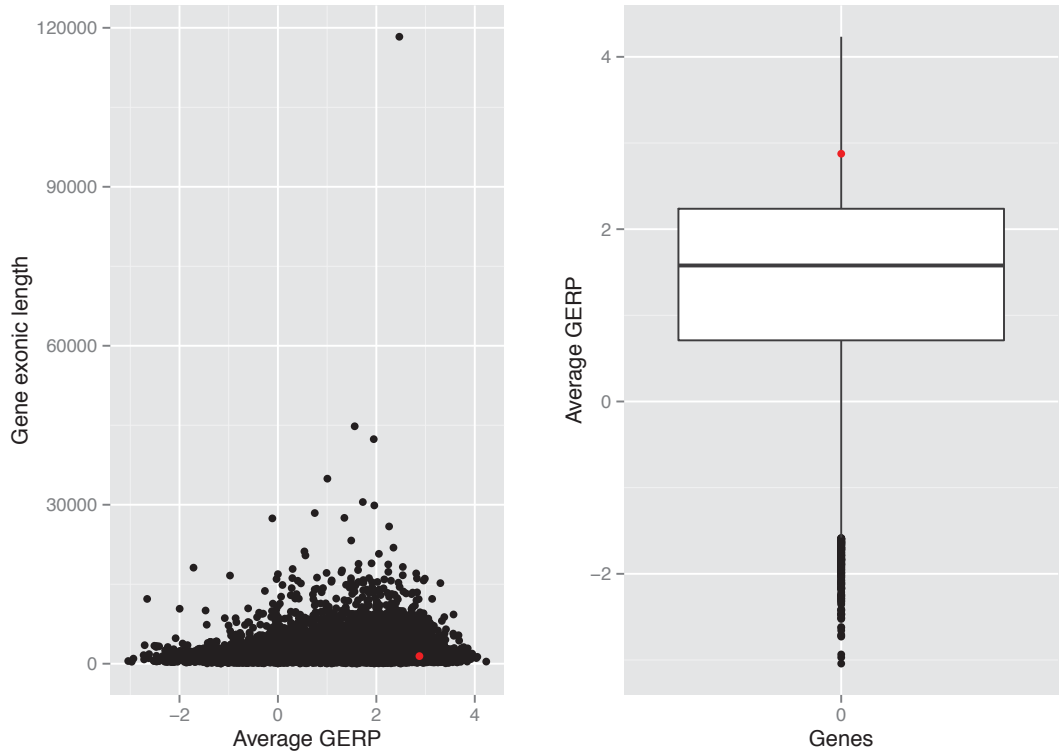


Figure 4-23 Average GERP scores averaged by gene length, NR2F2 denoted by the red color point (ranked 1059 out of 17,480 genes).

4.3.9 *NR2F2* rare coding variants in non-AVSD cases

There is considerable phenotypic heterogeneity in CHD whereby the same genes can be associated with diverse forms of CHD in humans e.g. *GATA4*, *NOTCH1, NKX2-5* and *CITED2*. Almost 45% of the CHD genes identified from mice knockouts have shown similarly diverse phenotypic outcomes [124, 504]. I therefore explored the frequency of *NR2F2* variants in other non-AVSD CHD cohorts available to us. With the help of our collaborators, we identified three additional CHD families with non-AVSD phenotypes with novel functional variants in *NR2F2*. In a patient with Tetralogy of Fallot (TOF) from the GO-CHD collection sequenced as part of the UK10K project, I detected a novel 3-bp insertion (p.Lys70LysGln). Using capillary sequencing, my colleague, Sarah Lindsay, was able to validate this variant and also to confirm it has been transmitted to two affected sons (one with AVSD and the other with aortic stenosis and ventricle septal defect) but not found in the healthy mother (Figure

4-25-a). In the second family from a Berlin CHD collection, and analyzed by both my colleague Marc-Phillip Hitz and myself, we found a trio of two healthy parents of an affected child with hypoplastic left heart syndrome (HLHS) and identified a *de novo* splice site (c.2359+1G>A) that was later confirmed by capillary sequencing by Sarah Lindsay, which is likely to cause skipping of the third exon (Figure 4-25-g). In addition to these two families, our collaborators David Wilson, and Catherine Mercer from the University of Southampton and David FitzPatrick from the University of Edinburgh were able to fine map a *de novo* balanced translocation 46,XY,t(14;15)(q23;q26.3) to the first intron of *NR2F2*, thus likely generating a null allele (Figure 4-24) by truncating the transcript after the first exon in a patient with coarctation of aorta (CoA).

Table 4-14 *NR2F2* sequence alterations identified in individuals with AVSD and other heart structural phenotypes.

Family	Subject	Sex	Phenotype	Mode of inheritance	cDNA position	Protein position	Amino Acid change	Variant type	GERP++
1	I:1	M	TOF	Unknown	208-211	70-71	K/KQ	In-frame insertion	-
1	II:1	M	cAVSD	Inherited	208-211	70-71	K/KQ	In-frame insertion	-
1	II:2	M	AS and VSD	Inherited	208-211	70-71	K/KQ	In-frame insertion	-
2	II:1	F	cAVSD	*De novo*	1022	341	S/Y	Missense	5.15
3	II:1	M	iAVSD	*De novo*	614	205	N/I	Missense	5.05
4	II:1	F	ubAVSD	Inherited	753	251	E/D	Missense	4.17
5	II:1	F	cAVSD	Inherited	1234	412	A/S	Missense	5.74
6	II:1	M	pAVSD	Unknown	509	170	D/V	Missense	5.00
7	II:1	F	HLHS	*De novo*	-	-	-	Splice donor	4.06
8	II:1	M	CoA	*De novo*	-	-	-	Balanced translocation	-

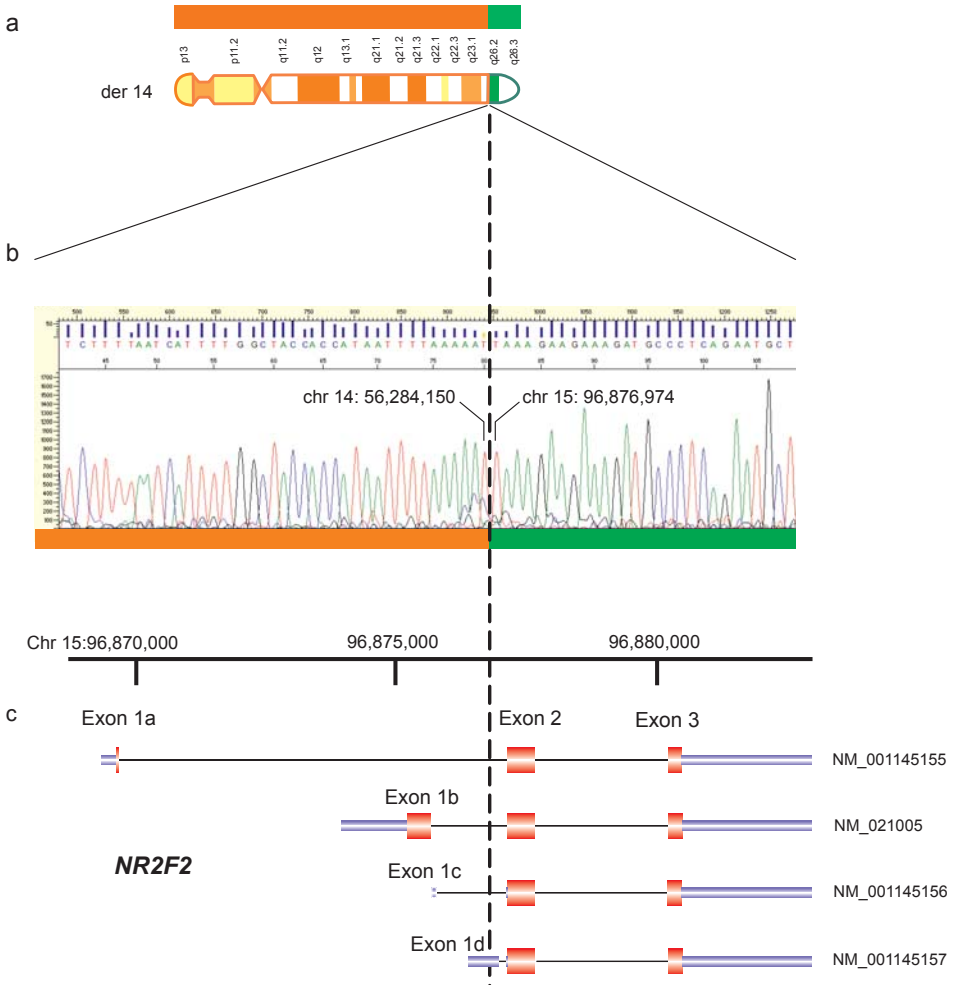


Figure 4-24 Derivative chromosome 14 breakpoint sequence. Ideogram of the derivative chromosome 14 (a) from patient with a balanced translocation [46,XY,t(14;15)(q23;q26.3)]. DNA sequence (b) of breakpoint junction between chromosome 14 and 15. Genomic organization of NR2F2 transcripts (c) and position of the breakpoint (figure courtesy of David Wilson and Catherine L. Mercer).

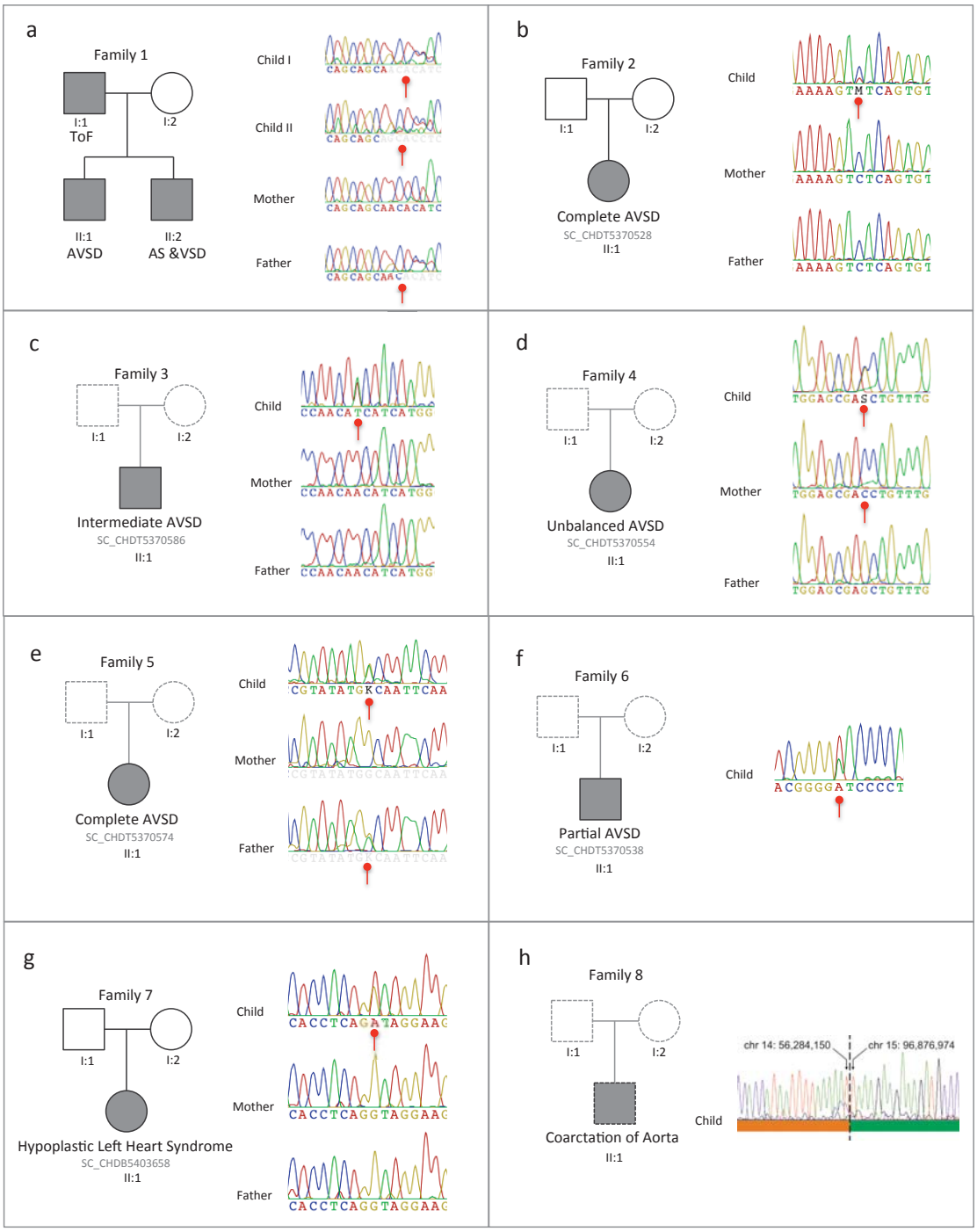


Figure 4-25: Pedigree charts and capillary sequencing results of *NR2F2* variants in eight CHD families. Solid lines in pedigree charts indicate both whole exome sequencing data and capillary sequencing are available while dash-line for samples with *NR2F2* capillary sequencing data only.

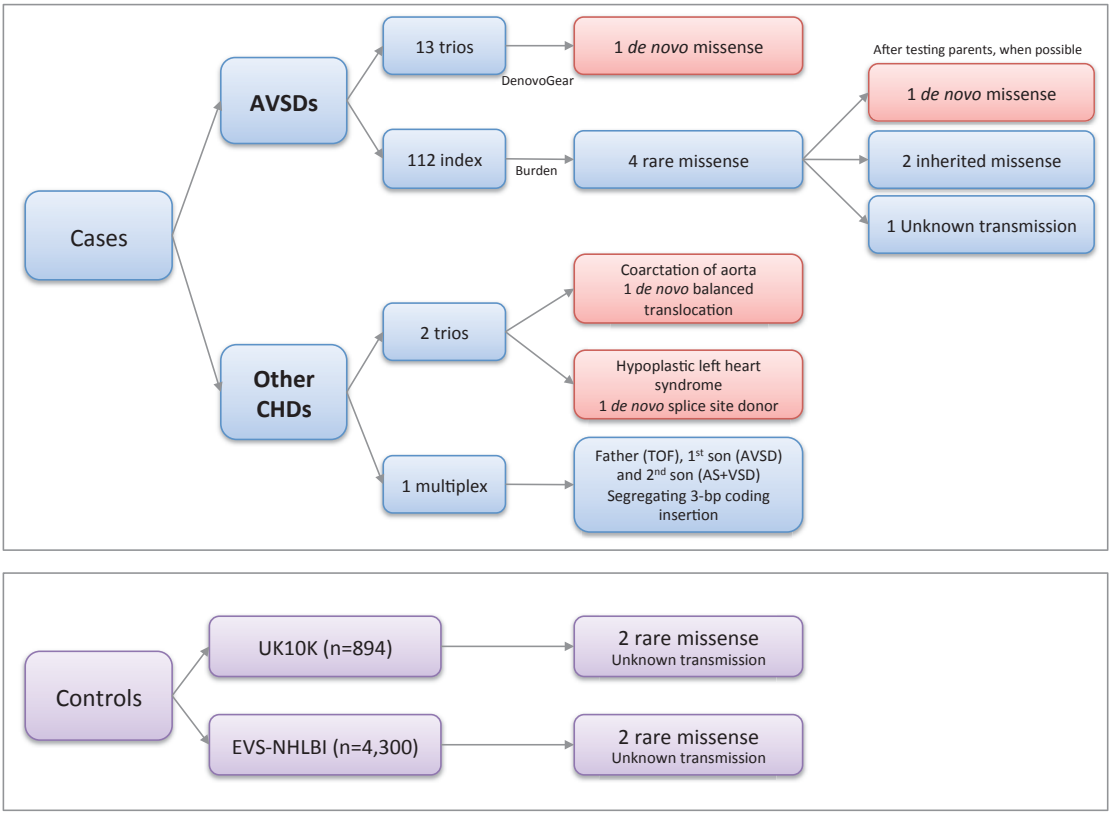


Figure 4-26 Number of cases and controls along with the number of NR2F2 variants and the mode of transmission in the discovery cohort. Red boxes are de novo variants. TOF: tetralogy of Fallot, AVSD: atrioventricular septal defects, AS: aortic stenosis, VSD: ventricular septal defect, CHDs: congenital heart defects.

4.3.10 *NR2F2* replication cohort

With the help of my colleagues, Sarah Lindsay at WTSI and Ashok Manickaraj at the SickKids hospital in Toronto, they were able to re-sequence the three coding exons in the major transcript of *NR2F2* in 248 additional AVSD samples, using PCR and capillary sequencing (Table 4-7), but they observed no additional rare functional variants in these samples. However, due to high GC content in the second *NR2F2* exon, the quality of capillary sequencing was not optimal despite many rounds of optimization. Other approaches such as targeted enrichment and sequencing on NGS platforms (see replication in chapter 3) or utilizing molecular inversion probe (MIP) [505] are potentially superior alternatives to capillary sequencing in any future follow up.

4.3.11 Family-based analysis using FEVA

To account for the rare Mendelian inherited variants, I used the FEVA software that I developed (described in chapter 2) to report a list of autosomal recessive candidate genes in the trios. Index cases were omitted in this analysis due to the lack of additional family information (e.g. paternal genotypes). Instead, I applied case/control analysis for the index cases (see next section).

The filters used by FEVA were aimed to capture rare coding variants assuming both parents were unaffected and complete penetrance. Table 2-11 lists the genotype combinations reported by FEVA under different inheritance models (see chapter 2 for details). The rare variants are defined based on a minor allele frequency < 1% in the 1000 genomes and 2,172 parental samples from the Deciphering Developmental Disorders (DDD) project. Coding variants were defined as any loss-of-function (e.g. frameshift, splice site donor or acceptor and stop gain and complex indels) or functional variants (e.g. missense and stop-loss).

This analysis identified 53 genes under different inheritance models (12 genes with homozygous variants, 31 genes with compound heterozygous and 10 genes on the X chromosome). Only one gene appears in more than one trio, *MADCAM1*, with the same homozygous frame-shift in two unrelated trios. *MADCAM1* gene encodes mucosal addressin cell-adhesion molecule-1 (MAdCAM-1) that is constitutively expressed in the gastrointestinal-associated lymphoid tissue. The knockdown mouse model [506] did not exhibit any structural phenotypes in the heart and thus this *MADCAM1* gene is unlikely to be involved in the AVSD phenotype. None of the other genes identified in FEVA output are known to cause CHD in human or in mouse models.

Table 4-15 The genotype combination in a complete trio reported by FEVA software under different models. Each trio includes an affected child (male or female) and two healthy parents. Each cell in the first column "genotype combinations" represents three genotypes in child, mother and father. "0" indicates a homozygous reference genotype, "1" is a heterozygous genotype, and "2" is a homozygous genotype in diploid chromosome (autosomal) or hemizygous in a haploid chromosome (e.g. X-chromosome in a male child). Y-chromosome and mitochondrial DNA are omitted from the table. Empty cells indicate that a given genotype combination is incompatible with Mendelian laws (e.g. 1,0,0 is *de novo*) or not expected under complete penetrance assumption (e.g. 1,1,1 is heterozygous in both the affected child and his parents). Only three genotype combinations were considered when I performed trios or multiplex analysis.

Genotype combinations	Autosomal	X- chromosome in an affected male child	X- chromosome in an affected female child
(1, 0, 0)			
(1, 0, 1)			
(1, 0, 2)			
(1, 1, 0)			
(1, 1, 1)			
(1, 1, 2)			
(1, 2, 0)			
(1, 2, 1)			
(1, 2, 2)			
(2, 0, 0)			
(2, 0, 1)			
(2, 0, 2)			
(2, 1, 0)		Hemizygous inherited from a carrier mother	
(2, 1, 1)	Homozygous in child and inherited from carrier parents		
(2, 1, 2)			
(2, 2, 0)			
(2, 2, 1)			
(2, 2, 2)			
(1,0,1) and (1,1,0)	Compound heterozygous in the child in a given gene		

4.3.12 Copy number variant (CNV) calling from exome data

Another class of variants known to increase the risk of isolated CHD is rare copy number variants (CNVs) [122]. I used CoNVex program [372], an algorithm developed by Parthiban Vijayarangakannan and Matthew Hurles, to detect copy number variation from exome and targeted-resequencing data using comparative read-depth. CoNVex corrects for technical variation between samples and detects CNV segments using a heuristic error-weighted score and the Smith-Waterman algorithm. The average number of called CNVs per sample is about 150-200 CNVs (both deletions and duplication). Since the false positive

rate (FPR) is generally high for most currently available methods that call CNV from the exome data, I used stringent filters to minimize the FPR. The first filter is the CoNVex score of 10 or more. This 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 confident calls. I also excluded common CNV, defined as CNV that appear in less than 1% of the population and appear in less than 5% (~20 samples) in the CHD exomes (i.e. internal control).

After applying these filters, I first looked for potential *de novo* CNV in the children and I detected four possible *de novo* duplications (Table 4-16). None of these genes appear to be expressed in the heart nor do they have any published knockout mouse models.

Table 4-16 Plausible de novo exome CNV in 13 AVSD trios

Sample id	Chr	Start	End	Size	Convex score	Type	Internal frequency	Genes
CHDL5262760	10	5201946	5202266	320	10.54	DUP	8	*AKR1CL1*
CHDL5262806	X	149012854	149014164	1,310	20.13	DUP	19	*MAGEA8*
CHDL5262830	12	9446101	9446662	561	10.67	DUP	16	*RP11-22B23.1*
CHDT5370568	9	15017219	15268088	250,869	17.68	DUP	1	*RP11-54D18.2, RP11-54D18.3, RP11-54D18.4, TTC39B, U6*

The next step was to look for the overlap between rare CNV and known CHD genes (400 genes), which yielded three rare duplications and one deletion in 125 AVSD cases (Table 4-17). Sample SC_CHDT5370541 carries a 150Kb long duplication on chromosome 21 and includes *RCAN1*, also known as Down syndrome critical region 1, *DSCR1* (Figure 4-27). This gene is a negative modulator of calcineurin/NFATc signaling pathway and expressed in embryonic brain and in the heart tube at E9.5-E10.5. The *DSCR1* expression in the heart has been detected in the truncus arteriosus, bulbus cordis and the primitive ventricle, which correlate with regions of endocardial cushion development and shown to be necessary for the normal development of heart valves [104, 462]. Moreover, the mice null model that lacks *NFATc1* expression dies secondary to

heart cushion defects [507]. The calcineurin/NFATc is known to regulate the Vascular Endothelial Growth Factor (*VEGF-A*), a known key regulator of endothelial cells. The *VEGF-A* levels need to be regulated precisely to ensure normal development of the heart cushions. Both over- and under- expression of the *VEFG-A* was shown to cause cushion development defects [508]. The presence of this small CNV may explain the AVSD phenotype observed in this patient. However, the burden of rare CNV overlapping this gene in CHD cases from the online Decipher database was not statistically significant when compared with healthy controls.

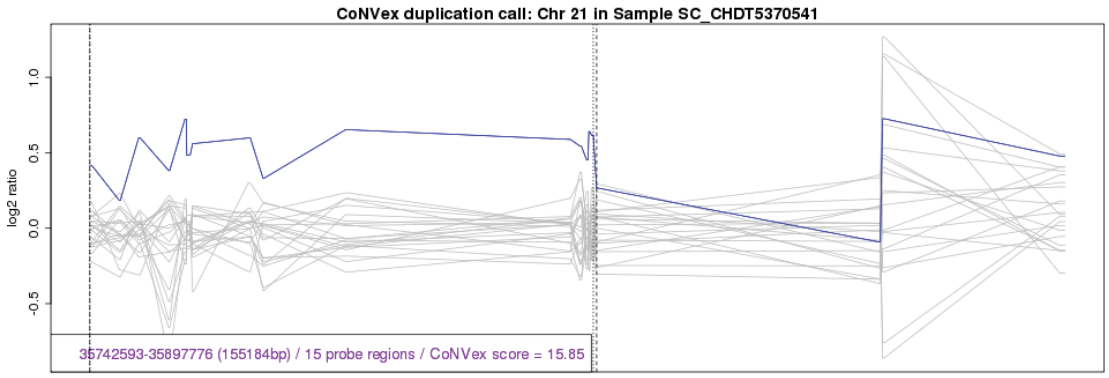


Figure 4-27 A 150 Kb duplication region detected on chromosome 21 and overlap with the critical region of Down syndrome (including *RCAN1* gene). The blue line is the log2 ratio in the patient (SC_CHDT5370541) with partial AVSD from SickKids hospital in Toronto collection. The grey lines log2ratio score for the same region in other CHD cases.

The only deletion I found overlapping with a known CHD gene is a 27 kb deletion that overlaps part of *EVC* and *CRMP1* genes (Figure 4-28). EVC is a known gene for Ellis-van Creveld Syndrome which is an autosomal recessive syndrome where patients exhibit disproportionate limb dwarfism, post-axial polydactyly, ectodermal dysplasia and congenital cardiovascular malformations in 60% of the patients of which the majority are AVSD [509]. However, the mouse model did not show a heart phenotype [510], *EVC* expression is detected in the secondary heart field, dorsal mesenchymal protrusion (DMP), mesenchymal structures of the atrial septum and the AV cushions [511]. Although the patient is not known to have Ellis-van Creveld syndrome, I searched the *EVC* gene for variants on the non-deleted allele (which may be hemizygous and appear to be homozygous, if they overlap the deletion) to see if the patient carries a combination of deletion

and a rare coding mutation (Table 4-18). I didn't find any known pathological mutation (HGMD version 2010.1) nor rare functional or loss of function variants. These findings suggest it is unlikely that the patient has Ellis-van Creveld Syndrome; but nonetheless this deletion may play a contributory role within an oligogenic framework.

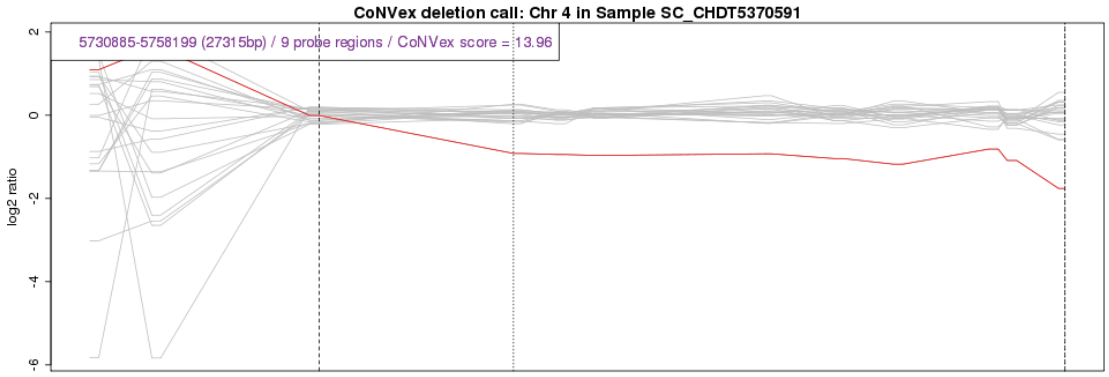


Figure 4-28 The log2ratio score of a 27 Kb deletion overlapping two genes, *EVC* and *CRMP1*. The grey lines log2ratio score for the same region in other CHD cases. The red line is the patient in which the variant was called.

Table 4-17 Rare CNV overlapping with known CHD genes

Sample id	Chr	Start	End	Size	Convex score	Type	Internal frequency	Genes
SC_CHDT5370524	1	100316428	100387368	70,940	25.07	DUP	1	***AGL***
SC_CHDT5370541	21	35742593	35897776	155,183	15.85	DUP	3	*AP000320.6, AP000322.53, AP000322.54, FAM165B,* ***KCNE1****,* ***KCNE2****,* ***RCAN1****, SNORA11*
SC_CHDT5370577	X	39921238	40586210	664,972	47	DUP	3	*ATP6AP2,* ***BCOR****, CXorf38, MED14, MPC1L, RP11-126D17.1, RP11-320G24.1, RP6-186E3.1, U7, Y_RNA, snoU13*
SC_CHDT5370591	4	5730885	5758199	27,314	13.96	DEL	1	*CRMP1,* ***EVC***

I also looked for rare coding variants under the dominant inheritance model overlapping with rare CNVs (i.e. possible compound heterozygous). I found nine rare CNVs with size ranges from 1 Kb to 2.5 Mb that overlap with at least one rare coding variant under the dominant model (i.e. inherited as a heterozygous from one parents). However, these CNVs were detected in many other CHD samples and also overlap with common CNV controls and hence are unlikely to be causal.

Table 4-18 List of variants called in EVC gene in sample (SC_CHDT5370591) with 27 Kb deletion detected by the exome CNV.

CHR	POS	REF	ALT	FILTER	Gene	Consequences	AF_MAX	Genotype	In deletion
4	5730954	G	A	PASS	*EVC*	INTRONIC	0.261155	HOM	Yes
4	5743509	C	T	PASS	*EVC*	SYNONYMOUS	0.998252	HOM	Yes
4	5743512	T	C	PASS	*EVC*	NON_SYNONYMOUS	0.947552	HOM	Yes
4	5747078	A	G	PASS	*EVC*	INTRONIC	0.699187	HOM	Yes
4	5747131	C	A	PASS	*EVC*	INTRONIC	0.611549	HOM	Yes
4	5750003	A	G	PASS	*EVC*	SYNONYMOUS	0.360892	HOM	Yes
4	5754544	T	C	PASS	*EVC*	INTRONIC	0.469816	HOM	Yes
4	5755542	C	A	PASS	*EVC*	NON_SYNONYMOUS	0.989837	HOM	Yes
4	5785442	G	A	PASS	*EVC*	NON_SYNONYMOUS	0.455801	HOM	Yes
4	5798627	G	A	PASS	*EVC*	INTRONIC	0.396341	HET	No
4	5800384	G	A	PASS	*EVC*	SYNONYMOUS	0	HET	No
4	5803669	T	C	PASS	*EVC*	SPLICE_SITE:INTRONIC	0.704724	HET	No
4	5803904	C	T	PASS	*EVC*	INTRONIC	0.704724	HET	No
4	5812195	A	G	PASS	*EVC*	INTRONIC	0.699187	HET	No
4	5812778	G	A	PASS	*EVC*	3PRIME_UTR	0.626016	HET	No

4.4 Discussion

AVSDs are an important subtype of CHD with a poorly understood genetic architecture. They represent 4-5% of all CHD and account for a large proportion of CHD in many syndromes such as Down and heterotaxy syndromes. The search for genetic causes in syndromic AVSD has been difficult. For example, the presence of three copies of chromosomes 21 increases the risk of AVSD but is not enough to explain why half of the Down syndrome patients do not exhibit other AVSD or other CHD. Many hypotheses have been suggested such as that a burden of rare missense in VEGF-A pathway genes (on chromosome 21) may play a role, but they are not conclusive [463]. On the other hand, it has been even more difficult to find the causative gene isolated non-syndromic AVSD cases. Only few studies were able to find plausible genetic causes in ~2% of the isolated AVSD cases on average in genes such as *CRELD1* and *GATA4*. In this chapter, I **combined exome data analysis** from hybrid family designs of 13 trios and 112 index cases to find genes enriched for rare coding variants (except silent variants).

What are the lessons from the burden analysis of rare coding variants in the case/control analysis?

There are many factors that could adversely affect a case/control analysis and should be addressed beforehand. These factors include sample contamination issues and population stratification. In this chapter I described two essential tests that removed ~11% of the control samples: the free-mix scores used to detect possible sample contamination and the principal component analysis (PCA) to detect possible population stratification. The free-mix scores were generated by 'verifyBAMid' software [488] by the UK10K team, which enabled me to remove ~8% (n=89 out of 1,008) of the UK10K neurological controls for possible contamination. Moreover, the PCA analysis worked very well and showed the relationship between the case/control samples in our exome projects to the four main populations from the HapMap project (CEU, YRI, CHB

and JPT) using ~10,000 common SNPs that are shared between them. This PCA analysis removed another ~3% of the controls (n=25) as possibly non-Caucasian samples.

Additionally, I observed another two factors with measurable effects that can be observed in the QQ plots of the case/control test results: the type of the pipelines used to call variants and the sample size of the cohort. The effect of the pipelines was observed when I evaluated different combinations of sample from both the GAPI and UK10K pipelines. Most of the QQ plots showed inflation (i.e. too many positive signals) when I used samples from two different pipelines. On the other hand, the QQ plots improved (showed less inflating) when I tested the variants in cases and controls called by the same pipeline. This is expected given what I already have learned from the comparisons of these pipelines (described in chapter 2), which showed that GAPI pipeline calls ~42% more rare missense variants than the UK10K pipeline. This can partially explain why I observed an inflated QQ plots when comparing AVSDs cases from GAPI pipeline with controls from the UK10K pipeline.

The second factor is the sample size of the cohort used in this analysis. QQ plots with small sample size < 100 showed a worse QQ inflation and improved dramatically when I increased the cases to ~260. These findings are also not surprising and I expect that increasing the sample size to a few more hundreds, possibly a few thousands, would be more appropriate sample size for this test.

What are the benefits of combining the *de novo* analysis with the case/control?

Although the burden analysis of rare missense variants has identified *NR2F2* as one of the enriched genes for rare missense variants in the cases, the *NR2F2* gene was not the top candidate gene and it did not reach a genome-wide statistical significance. This case/control analysis identified five AVSD cases and two controls with rare missense variants (fisher exact test, *P*= 0.00011, when considering AVSD cases from GAPI pipeline only). This modest result led me to

overlook *NR2F2* gene initially. Only when I performed the *de novo* analysis and found that one of the five rare missense variants in AVSD cases was actually a *de novo* variant, that this gene made it back to the top of the AVSD candidate gene list.

This shows that even when the sample size of this AVSD cohort is underpowered for the case/control analysis, intersecting gene lists from both *de novo* and case/control analyses can salvage the latter.

How *NR2F2* mutations cause the congenital heart defects?

NR2F2 belongs to a small family of the steroid/thyroid hormone receptor nuclear superfamily which includes two related but distinct genes: *NR2F1* (or *COUP-TFI*) and *NR2F2* (or *COUP-TFII*). Both genes are involved in many cellular and developmental processes. While *NR2F1* is mainly involved in neural development, *NR2F2* is expressed and involved in the organogenesis of the stomach, limbs, skeletal muscles and the heart (reviewed in ref [512]). The ligand for NR2F2 is not yet known. The missense variants seen in patients are distributed throughout NR2F2, with three falling in the ligand-binding domain (p.Asn205Ile, p.Glu251Asp and p.Ser341Tyr) of which two can be mapped to a previously determined partial crystal structure for this domain [502] (Figure 4-20 d-f): p.Asn205Ile is expected to perturb ligand binding whereas p.Ser341Tyr is predicted to destabilize the homodimerization domain.

The *Nr2f2* mouse null model leads to embryonic lethality with severe hemorrhage and failure of the atria and sinus venosus to develop past the primitive tube stage [513]. A more recent hypomorphic *Nr2f2* mouse mutant exhibits a more specific heart phenotype with atrioventricular septal and valvular defects due to the disruption of endocardial cushion development in a dosage-sensitive fashion. This is partially driven by defective endothelial-mesenchymal transformation (EMT) and the hypocellularity of the atrioventricular canal accompanied by down regulation of *Snai1* [501]. Our knockdown and over-expression studies of *nr2f2* in zebrafish confirmed that the

developing vertebrate embryo is exquisitely sensitive to *nr2f2* dosage (data not shown), such that knockdown rescue experiments are precluded.

In addition to the direct role of *NR2F2* mutations in causing congenital heart defects, given its dosage sensitivity, *NR2F2* may potentially also act as an environmentally responsive factor by mediating the effect of known non-genetic CHD risk factors such as high glucose [514] and retinoic acid levels [515]. Insulin and glucose levels are known to negatively control *NR2F2* expression via the Foxo1 pathway in hepatocyte and pancreatic cells [516]. Furthermore, *NR2F2* has been shown to play a critical role in retinoic acid signaling during development [517]. Further investigations are needed to determine how glucose and retinoic acid levels may alter *NR2F2* expression in the developing heart.

Is there a genotype-phenotype correlation between the coding variants in *NR2F2* and the CHD subtypes?

In addition to the five AVSD families with rare missense variants in *NR2F2* gene (two arose *de novo*, two were inherited and one unknown inheritance), with the help of my collaborators, we found three non-AVSD families with rare inherited or *de novo* variants in *NR2F2*. The first was a novel coding 3bp insertion (p.Lys70LysGln) in a parent with Tetralogy of Fallot that also co-segregate in two affected sons (one with AVSD and one with aortic stenosis and ventricle septal defect). The second variant was a *de novo* balanced translocation 46,XY,t(14;15)(q23;q26.3) at the first intron of *NR2F2* in a patient with coarctation of aorta. The third variant was a *de novo* splice site (c.2359+1G>A) that is likely to skip the third exon which later was seen in a child with hypoplastic left heart syndrome (Table 4-14, Figure 4-25 and Figure 4-26).

Moreover, a previous case report of a child with a terminal deletion of 15q and septal defects (VSD and ASD) proposed *NR2F2* as a candidate gene for CHD as it falls within a critical interval deleted in the subset of patients that have CHD in addition to the canonical syndromic features [518]. Based on a literature survey

of rare variants overlapping *NR2F2* gene in human (carried out by Dr. Catherine Mercer, personal communication) Dr. Matthew Hurles and myself compared the cardiac phenotypes of thirteen patients with loss-of-function variants (including published whole gene deletions) and eight patients with coding sequence variants revealed an intriguing genotype-phenotype correlation. Most patients with loss-of-function variants had Left Ventricular Outflow Tract Obstruction (LVOTO, N=9), but none had AVSD, although most (N=8) had ASD or VSD. Conversely, six out of eight patients with coding sequence variants had AVSD, but only one had LVOTO and one had VSD. This observation that the more severe mutations result in LVOTO in addition to septal defects merits further investigation in larger numbers of patients with *NR2F2* mutations.

Does the negative result in the replication study suggest a 'winner's-curse'?

The number of rare missense variants I observed in the *NR2F2* gene from controls was extremely rare (only ~0.0009% based on the analysis of more than 10,000 samples from different internal and external whole genome/exome sequencing projects). On the other hand, the analysis of the primary AVSD cohort (n=125) identified five patients with either rare inherited or *de novo* missense variants in the *NR2F2* gene (4%). This is percentage is unusually high when compared with candidate re-sequencing studies in CHD where the average number of patients detected with rare coding variants is usually around ~2%. Hence, it was surprising that the replication study of 245 AVSD cases has not identified a single case with rare missense variant in the *NR2F2* gene.

One important explanation for the negative results in the replication experiment is the winner's curse, a well-known phenomenon in the world of genome-wide associations studies [519]. This phenomenon is an ascertainment bias that leads overestimating the penetrance and allele-frequency parameters for the associated variant, which usually lead to negative results in the subsequent results. Did I underestimate the number of samples required for the replication study in isolated AVSDs? Most likely.

Another factor that to the negative results is the difference between the sequencing methods used to screen *NR2F2* gene for rare coding variants in the primary and replication cohort. My collaborators (Dr. Sarah Lindsay at the Wellcome Trust Sanger Institute and Ashok Kumar at the University of Toronto) have used capillary sequencing to screen the *NR2F2*'s three exons. They both have reported difficulties in the *NR2F2* sequencing due to high GC content resulted in a high failure rate of sequencing experiments. The is unlike the exome sequence data, which showed very good sequence coverage of *NR2F2* exons and all coding variant detected in the cases were confirmed to be true positive. This suggests that we might have missed true missense variant(s) by using the capillary sequencing in such difficult regions and an alternative screening methods (such as custom designs baits or MIP coupled with NGS) is a better alternative approach for the next replication study in *NR2F2*.

Are there other AVSDs candidate genes found in this cohort?

The family-based analysis (FEVA) analysis of rare recessive variants did not identify any strong AVSD candidate gene, which is not unexpected given the small number of trios included in this cohort (n=13). The CNV analysis based on exome data identified few interesting variants such as a 27kb deletion that overlaps with *EVC* gene, a known gene for the Ellis-van Creveld syndrome where CHD occur in ~60% and most are AVSD. Although Ellis-van Creveld syndrome is known to be a recessive syndrome, there are examples of hypomorhpic mutations in the *EVC* gene that are found to cause a phenotype of cardiac and limb defects that is less severe than typical Ellis-van Creveld syndrome [520]. However, this deletion needs to be confirmed using an independent method (MLPA or array CGH) before considering it any further.

Future directions

Increasing the sample size of the replication cohort and also including non-AVSD cases are likely to essential for future *NR2F2* replication studies in order to

understand the involvement of this gene's mutations in various CHD subtypes. The two study designs used in this chapter, the trios and the case/control, showed very promising results and using them in future isolated AVSD studies, whether in combination or separately, is expected to lead to the discovery of other genes. More importantly, calling the exome variants across all samples by the same pipeline is strongly advised to avoid spurious false positive findings introduced by the subtle differences in filters thresholds and various other components of the calling pipelines.

In summary, these findings add *NR2F2* to the short list of dosage-sensitive regulators such as *TBX5*, *TBX1*, *NKX2-5* and *GATA4* that have been shown, when mutated, to interfere with normal heart development and that lead to the formation of CHD in both mice and humans. By virtue of their dosage sensitivity, these master regulators potentially play a key role in integrating genetic and environmental risk factors for abnormal cardiac development.

5 | Discussion

In this thesis I explored different subtypes of congenital heart defect (CHD) using next-generation sequencing (NGS) data with a focus on family-based study designs such as parent-offspring trios. Even with the relatively small sample sizes of the cohorts studied in this thesis, I was able to detect three clearly pathogenic genes: *NOTCH1* and *JAG1* in isolated tetralogy of Fallot and *NR2F2* in isolated atrioventricular septal defects.

What did I learn about exome analysis pipelines?

At the beginning of my PhD studies, variant calling from whole genome or whole exome sequencing data was still in its infancy. It was not clear what were the best practices, pipelines, tools or filtering strategies required to achieve high levels of sensitivity and specificity for variant identification. This led me to investigate different aspects of the variant calling workflow to determine appropriate callers and filters to achieve high specificity and sensitivity.

Initially, I assessed **sequence and variant calling parameters** such as phred-like quality (QUAL), strand bias (SB), quality-by-depth (QD) and genotype quality (GQ) in order to set thresholds to eliminate low quality variants. These filters and thresholds worked well for the early sample releases, but as the underlying probabilistic models for calling and filtering variants improved, these filters changed accordingly and they will probably continue to change in the foreseeable future. Newer parameters of sequence data and variant calling have emerged and they are replacing many previous filtering strategies (for example the Variant Quality Score Recalibration (VQSR) filter from GATK caller has been suggested as a superior quality filter for single nucleotide variants from exome sequencing, but not indels). Currently, choosing the right set of filters and

thresholds is an area that needs to be revisited on a regular basis in order to adhere to the best practices available.

Another important part of variant calling workflows, which is usually overlooked, is **how to merge variants identified by two or more callers** (e.g. Samtools and GATK). If the two callers disagree on an alternative allele or a genotype, which caller should be used as the default? When I started my projects I decided, naively, to use GATK as the default caller over both Samtools and Dindel, for samples called by the Genome Analysis Production Informatics (GAPI) pipeline. However, when I investigated this issue in more detail later on, I discovered a complex relationship between the type of the caller used as a default caller, and the number and type of rare coding variants identified and reported for downstream analysis. For example, Samtools tends to call more rare loss of function variants (~8 per sample on average) that are either missed by GATK or have been flagged by GATK as low quality variants. I was able to show that these variants exhibit a low transition/transversion ratio, which is indeed a sign of being low quality variants. In studies with a small number of samples this might not be a major issue, but for large-scale projects with hundreds or thousands of samples such as the Deciphering Developmental Disorders project with 12,000 affected children, this can have a huge effect on the amount of downstream work required for validation and / or functional experiments. These findings hold true for the version of the callers used to call variants in my samples, but it is expected to change when using a different version of the same caller, and thus it is important to perform this detailed analysis whenever a newer version of a variant caller is implemented.

Small decisions such as what threshold of a filter should be used, or which is the default variant caller, can lead to big differences in the type, number and quality of the variants identified in whole exome data, especially the rare coding variants of greatest interest in rare disease studies. This was clearly manifested by the variant differences I identified between **two analytical pipelines** that were used to call variants from the CHD samples described in this thesis: Genome Analysis Production Informatics (GAPI) and UK10K. Both pipelines used different

numbers and versions of the variant callers and they also adopted variable filters and thresholds. Each difference might have a small effect on its own, but their cumulative effects are appreciable. The most obvious differences I observed were in the number of rare coding variants in the GAPI pipeline which called (~42%) rare missense variants and almost 4.4-fold more coding insertion/deletion (indels) than the UK10K pipeline. When samples are used from both pipelines, as they were in the burden analysis of rare missense variants in **chapter 4**, I noticed an inflation of quantile-quantile (Q-Q) plots. An obvious explanation was that the inflation was caused by the high number of rare missense variants in the GAPI pipeline compared with the controls from the UK10K pipeline. However, it is likely that the explanation is probably more complex, and is caused by multiple factors. More work is required to investigate the origin of these differences.

What did I learn about tetralogy of Fallot?

The **two-stage study design** I used to investigate the genetic architecture of isolated **tetralogy of Fallot** enabled me to detect two clearly pathogenic genes: *NOTCH1* and its ligand *JAG1* in a cohort of 238 parent-offspring trios. Although both genes have been associated with congenital heart defects in the past, their involvement in the isolated tetralogy of Fallot is less well appreciated. Rare coding variants in *NOTCH1* have been linked to familial forms of left ventricular outflow tract malformations more often than with the malformations of the right side of the heart. Similarly, mutations in *JAG1* are usually associated with Alagille syndrome where CHD occurs in ~90% of the patients (6-17% are ToF) more often than with non-syndromic tetralogy of Fallot. I was able to detect *de novo* coding variants (except silent variants) in these genes in 2.5% of patients in this cohort. These variants included four *de novo* coding variants in the *NOTCH1* gene and two *de novo* coding variants in the *JAG1* gene. Interestingly, two-thirds of these *de novo* variants are loss-of-function, which showed up as a highly statistically significant burden of *de novo* loss-of-function in the *NOTCH1* gene (P=3.8 ×10^{-9}).

More interestingly, a **theme has emerged when I combined** *de novo* variant analysis with other analyses that target rare coding variants with presumably intermediate effect size (i.e. incomplete penetrance). I identified two genes, *NOTCH1* and *ARHGAP35*, both with *de novo* functional or loss-of-function variants, and both were also enriched for rare inherited missense variants. The case/control analysis identified *NOTCH1* as being enriched for rare missense variants (P=8.8 × 10^{-05}). On the other hand, the modified transmission disequilibrium test (TDT) identified an over-transmission of rare missense variants in the *ARHGAP35* (P=0.02).

Collectively, these genes have five *de novo* variants where all but one, are loss-of-function variants. This observation suggests that two classes of variants contribute to the isolated tetralogy of Fallot. The first group is rare coding variants with large effect size, mainly loss-of-function, that are able to cause the phenotype when they occur *de novo*. The second group is rare, typically missense, variants that increase the risk of isolated tetralogy of Fallot but are not sufficient to cause the phenotype by themselves. This group might require additional in *cis*- or *trans*- variants in order to cause the phenotype. One way to investigate this possibility is the digenic inheritance model that I described in **chapter 3**. Although the digenic inheritance analysis has identified a few interesting gene pairs such as *ZFPM2-CTBP2* that are enriched for rare missense variants in cases compared with 1,080 controls, the sample size is clearly underpowered, so I was not able to obtain signals that are statistically significant at the genome-wide level.

What did I learn about isolated atrioventricular septal defects?

Similarly, combining *de novo* analysis with case/control analysis enabled me to identify *NR2F2* as a novel candidate gene **for isolated atrioventricular septal defects** (AVSD) in human (**chapter 4**). Although the case/control analysis of a burden of rare missense variants burden did not, on its own, identify *NR2F2* as the most significant gene, it was the subsequent *de novo* analysis that identified this gene as the most intriguing candidate gene in this cohort.

NR2F2 is one of the most conserved genes across the genome and exhibits very little variation in populations, which supports its fundamental roles in the development of many organs, including the heart. Additionally, the published conditional knockout mouse model recapitulated many of the atrioventricular septal defects observed in human. These findings have been shown by others to be driven by defective endothelial-mesenchymal transformation (EMT) and the hypocellularity of the atrioventricular canal, accompanied by down regulation of the *Snai1* gene. Moreover, the results from luciferase assays (appendix B) performed by my colleague, Sebastian Gerety, indicate that all Nr2f2 coding sequence variants identified from the AVSD cohort had a measurable impact on transcriptional activation in at least one target gene. Further modelling work will be required to clarify whether these differences between target genes translate into distinct biological mechanisms of disease, affecting single or multiple molecular interactions required for heart morphogenesis.

Expanding the search for *NR2F2'*s mutations in other CHD subtypes revealed its involvement in tetralogy of Fallot, hypoplastic left heart syndrome and coarctation of the aorta. This analysis increased the total number of CHD families with *NR2F2* to eight (I have identified six CHD families while the other two CHD families were identified by my collaborators: David Wilson, David FitzPatrick and Catherine Mercer who identified a *de novo* balanced translocation in a child with coarctation of the aorta and Marc-Phillip Hitz who identified a *de novo* splice site in a child with hypoplastic left heart syndrome). These findings suggest *NR2F2* as a **novel dosage-sensitive regulator gene** involved in the CHD in human similar to other well-known CHD genes such as *TBX5*, *TBX1*, *NKX2-5* and *GATA4*. I hypothesise that these master regulators potentially play a key role in integrating genetic and environmental risk factors for abnormal cardiac development, although testing this hypothesis will require substantial downstream work.

What did I learn about study designs?

The two most **informative study designs** I evaluated in my thesis are the trio-based and the case/control designs. The trio family-based design is a versatile design since it is amenable to different analyses aimed to investigate rare coding variants with large size effect as well as variants with intermediate effect sizes. *De novo* analysis is the main test used to investigate variants with large effect size. Less commonly used, the modified transmission disequilibrium test (TDT) tries to identify over-transmission of rare variants from healthy parents to their affected children, as well as the digenic inheritance analysis which targets rare variants in affected children inherited from two different parents. The case/control analysis worked surprisingly well given the small size of the cohorts in this thesis. Its success is most likely attributed to being used in combination with the results from the *de novo* analysis. Nonetheless, performing case/control analysis in larger sample size of homogenous CHD cohorts is expected to identify additional genes involved in congenital heart defects. Other study designs I used such as affected parent-child and affected sib-pairs were not as successful, but this is likely to be due to the small sample size of these studies, and the difficulty in identifying additional families with similar mutations.

What were the limitations of my work?

Next-generation sequence (NGS) platforms have revolutionized the way we identify causal genes in monogenic disorders. This technology has helped me to identify different causal genes in two non-syndromic CHD subtypes. Nonetheless, NGS platforms impose some major **analytical challenges**. The most important one is the fact that my analysis, in common with all such analyses, has identified too many variants of unknown significance (VUS). This reflects our current state of very limited understanding of the function of most genes and the consequences of most variants. One way to overcome this problem in gene discovery analysis, will be to increase the sample size in order to increase the power of genetic analyses. International collaborations and data sharing will be important for increasing sample sizes. For VUS in known CHD genes, functional

assays *in vivo* or *in vitro* may help to confirm their pathogenicity, although even these assays will have their associated false positives and false negatives.

How do my findings relate to other peoples work?

Recently, Zaidi *et al.* used NGS to sequence the whole exome in a trio cohort of 362 severe cases of syndromic and non-syndromic CHD and predicted that *de novo* point mutations in several hundreds of genes may contribute to ~10% of severe CHD cases [256]. This estimation is difficult to ascertain using the samples described in my thesis, since I have a much smaller sample size of trios (n=43 complete trios with whole exome sequence data). Nonetheless, I was able to identify likely pathogenic *de novo* variants in *NOTCH1* and *JAG1* in 2.5% of isolated tetralogy of Fallot (six out of 238 trios) and about ~12% in atrioventricular septal defects trios (two out of 16 complete trios that were available with either exome data or capillary sequencing) but given the other candidate genes that I identified with *de novo* variants (e.g. *ZMYM2, ARHGAP35, HDAC3*). My results are broadly consistent with the conclusion by Zaidi et al.

Future directions

Selecting an optimal variant calling pipeline is not an easy task and once one is implemented, any potential upgrade or new pipeline needs to be assessed in considerable detail to ensure that data quality is improved. Equally importantly, using a single, consistent, pipeline is essential in order to obtain consistent datasets, which helps to avoid complicating any downstream analyses.

Future CHD studies will require **larger sample sizes,** possibly of the order of a few thousand samples, in order to achieve enough power to identify a substantial fraction of recurrently mutated causal genes. Given the rarity of many CHD subtypes, a **national and international network of collaborators** is necessary to collect enough samples for parent-offspring complete trios and/ or case-control designs, both of which have been shown to be suitable study designs for isolated CHD.

Beside the genetic components required to support newly identified CHD genes in trios and case/control study designs, **functional experiments** are essential to confirm the pathogenic effect of genes in animal models using knockout or knockdown experiments in mouse and zebrafish models. Where appropriate, the pathogenic effect of specific variants can also be investigated using cell-based assays such as luciferase activity experiments. Moreover, integrating exome and genome sequence data with gene expression data using RNA-Seq from fetal heart tissues at different developmental stages are likely to be a helpful tool to prioritize candidate genes. Integrating high-throughput genetics, functional genomics and cellular and animal modeling will require concerted effort and collaboration.

References

1. Pollak, K. and E.A. Underwood, *The healers: the doctor, then and now*. 1968, London,: Nelson. x, 246 p.
2. Neill, C.A. and E.B. Clark, *Tetralogy of Fallot. The first 300 years.* Tex Heart Inst J, 1994. **21**(4): p. 272-9.
3. Maheshwari S, K.V., *Textbook of Cardiology (A Clinical & Historical Perspective)* N.C.N. H K Chopra Editor. 2012, Jaypee Brothers Medical Publishers. p. 270-282.
4. Robert E. Gross, M.D.J.P.H., M.D., *Surgical ligation of a patent ductus arteriosus report of first successful case.* JAMA, 1939. **112**(8): p. 729-731.
5. Rashkind, W.j., *Pediatric Cardiology: A Brief Historical Perspective* . Pediatr Cardiology, 1979. **1**: p. 63-71.
6. Baars, H.F., J.J.v.d. Smagt, and P.A. Doevendans, *Clinical cardiogenetics*. 2011, London: Springer. xv, 455 p.
7. Olley, P.M., F. Coceani, and E. Bodach, *E-type prostaglandins: a new emergency therapy for certain cyanotic congenital heart malformations.* Circulation, 1976. **53**(4): p. 728-31.
8. Laurenceau, J.L., et al., *[Study of tetralogy of Fallot by echocardiography].* Arch Mal Coeur Vaiss, 1975. **68**(5): p. 505-12.
9. Ferencz, C.L., CA, Correa-Villasenor, Wilson,PD, *Genetic and Environmental Risk Factors of Major Cardiovascular Malformations, The Baltimore-Washington Infant Study, (1981-1989)*. 1997: Perspectives in Pediatric Cardiology, vol.5. Armonk, N.Y: Futura Publishing Co.Inc.
10. !!! INVALID CITATION !!!
11. Gupta, V. and K.D. Poss, *Clonally dominant cardiomyocytes direct heart morphogenesis.* Nature, 2012. **484**(7395): p. 479-84.
12. Yelon, D., *Developmental biology: Heart under construction.* Nature, 2012. **484**(7395): p. 459-60.
13. van der Linde, D., et al., *Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis.* J Am Coll Cardiol, 2011. **58**(21): p. 2241-7.
14. Fahed, A.C., et al., *Genetics of congenital heart disease: the glass half empty.* Circ Res, 2013. **112**(4): p. 707-20.
15. van der Bom, T., et al., *The changing epidemiology of congenital heart disease.* Nat Rev Cardiol, 2011. **8**(1): p. 50-60.
16. Brickner, M.E., L.D. Hillis, and R.A. Lange, *Congenital heart disease in adults. First of two parts.* N Engl J Med, 2000. **342**(4): p. 256-63.
17. Soulvie, M.A., et al., *Psychological Distress Experienced by Parents of Young Children With Congenital Heart Defects: A Comprehensive Review of Literature.* Journal of Social Service Research, 2012. **38**(4): p. 484-502.
18. Sadoh, W.E., D.U. Nwaneri, and A.C. Owobu, *The cost of out-patient management of chronic heart failure in children with congenital heart disease.* Niger J Clin Pract, 2011. **14**(1): p. 65-9.
19. Hoffman, J.I. and S. Kaplan, *The incidence of congenital heart disease.* J Am Coll Cardiol, 2002. **39**(12): p. 1890-900.

20. Bernier, P.L., et al., *The challenge of congenital heart disease worldwide: epidemiologic and demographic facts.* Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu, 2010. **13**(1): p. 26-34.
21. Warnes, C.A., et al., *Task force 1: the changing profile of congenital heart disease in adult life.* J Am Coll Cardiol, 2001. **37**(5): p. 1170-5.
22. Marelli, A.J., et al., *Congenital heart disease in the general population: changing prevalence and age distribution.* Circulation, 2007. **115**(2): p. 163-72.
23. van der Velde, E.T., et al., *CONCOR, an initiative towards a national registry and DNA-bank of patients with congenital heart disease in the Netherlands: rationale, design, and first results.* Eur J Epidemiol, 2005. **20**(6): p. 549-57.
24. Nora, J.J. and A.H. Nora, *Maternal transmission of congenital heart diseases: new recurrence risk figures and the questions of cytoplasmic inheritance and vulnerability to teratogens.* Am J Cardiol, 1987. **59**(5): p. 459-63.
25. Nora, J.J. and A.H. Nora, *The evolution of specific genetic and environmental counseling in congenital heart diseases.* Circulation, 1978. **57**(2): p. 205-13.
26. Nora, J.J. and A.H. Nora, *Recurrence risks in children having one parent with a congenital heart disease.* Circulation, 1976. **53**(4): p. 701-2.
27. Nora, J.J., C.W. McGill, and D.G. McNamara, *Empiric recurrence risks in common and uncommon congenital heart lesions.* Teratology, 1970. **3**(4): p. 325-30.
28. Nora, J.J., *Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction.* Circulation, 1968. **38**(3): p. 604-17.
29. Burn, J., et al., *Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study.* Lancet, 1998. **351**(9099): p. 311-6.
30. Burn, J. and G. Corney, *Congenital heart defects and twinning.* Acta Genet Med Gemellol (Roma), 1984. **33**(1): p. 61-9.
31. Fesslova, V., et al., *Recurrence of congenital heart disease in cases with familial risk screened prenatally by echocardiography.* J Pregnancy, 2011. **2011**: p. 368067.
32. Oyen, N., et al., *Recurrence of congenital heart defects in families.* Circulation, 2009. **120**(4): p. 295-301.
33. Hardin, J., et al., *Increased prevalence of cardiovascular defects among 56,709 California twin pairs.* Am J Med Genet A, 2009. **149A**(5): p. 877-86.
34. Lewin, M.B., et al., *Echocardiographic evaluation of asymptomatic parental and sibling cardiovascular anomalies associated with congenital left ventricular outflow tract lesions.* Pediatrics, 2004. **114**(3): p. 691-6.
35. Gill, H.K., et al., *Patterns of recurrence of congenital heart disease: an analysis of 6,640 consecutive pregnancies evaluated by detailed fetal echocardiography.* J Am Coll Cardiol, 2003. **42**(5): p. 923-9.
36. Shieh, J.T. and D. Srivastava, *Heart malformation: what are the chances it could happen again?* Circulation, 2009. **120**(4): p. 269-71.
37. Oyen, N., et al., *Recurrence of discordant congenital heart defects in families.* Circ Cardiovasc Genet, 2010. **3**(2): p. 122-8.
38. Hoffman, J.I., *Incidence of congenital heart disease: II. Prenatal incidence.* Pediatr Cardiol, 1995. **16**(4): p. 155-65.

39. Blue, G.M., et al., *Congenital heart disease: current knowledge about causes and inheritance.* Med J Aust, 2012. **197**(3): p. 155-9.
40. Talner, C.N., *Report of the New England Regional Infant Cardiac Program, by Donald C. Fyler, MD, Pediatrics, 1980;65(suppl):375-461.* Pediatrics, 1998. **102**(1 Pt 2): p. 258-9.
41. Chang, R.K., M. Gurvitz, and S. Rodriguez, *Missed diagnosis of critical congenital heart disease.* Arch Pediatr Adolesc Med, 2008. **162**(10): p. 969-74.
42. Hoffman, J.I., *It is time for routine neonatal screening by pulse oximetry.* Neonatology, 2011. **99**(1): p. 1-9.
43. Kemper, A.R., et al., *Strategies for implementing screening for critical congenital heart disease.* Pediatrics, 2011. **128**(5): p. e1259-67.
44. de-Wahl Granelli, A., et al., *Impact of pulse oximetry screening on the detection of duct dependent congenital heart disease: a Swedish prospective screening study in 39,821 newborns.* BMJ, 2009. **338**: p. a3037.
45. Chaturvedi, V. and A. Saxena, *Heart failure in children: clinical aspect and management.* Indian J Pediatr, 2009. **76**(2): p. 195-205.
46. Verheugt, C.L., et al., *Gender and outcome in adult congenital heart disease.* Circulation, 2008. **118**(1): p. 26-32.
47. Verheugt, C.L., et al., *Long-term prognosis of congenital heart defects: a systematic review.* Int J Cardiol, 2008. **131**(1): p. 25-32.
48. Walsh, E.P. and F. Cecchin, *Arrhythmias in adult patients with congenital heart disease.* Circulation, 2007. **115**(4): p. 534-45.
49. Rhodes, L.A., et al., *Arrhythmias and intracardiac conduction after the arterial switch operation.* J Thorac Cardiovasc Surg, 1995. **109**(2): p. 303-10.
50. van den Bosch, A.E., et al., *Long-term outcome and quality of life in adult patients after the Fontan operation.* Am J Cardiol, 2004. **93**(9): p. 1141-5.
51. Tleyjeh, I.M., et al., *Temporal trends in infective endocarditis: a population-based study in Olmsted County, Minnesota.* JAMA, 2005. **293**(24): p. 3022-8.
52. Niwa, K., et al., *Infective endocarditis in congenital heart disease: Japanese national collaboration study.* Heart, 2005. **91**(6): p. 795-800.
53. Di Filippo, S., et al., *Current patterns of infective endocarditis in congenital heart disease.* Heart, 2006. **92**(10): p. 1490-5.
54. Duffels, M.G., et al., *Pulmonary arterial hypertension in congenital heart disease: an epidemiologic perspective from a Dutch registry.* Int J Cardiol, 2007. **120**(2): p. 198-204.
55. Diller, G.P. and M.A. Gatzoulis, *Pulmonary vascular disease in adults with congenital heart disease.* Circulation, 2007. **115**(8): p. 1039-50.
56. Barst, R.J., et al., *Diagnosis and differential assessment of pulmonary arterial hypertension.* J Am Coll Cardiol, 2004. **43**(12 Suppl S): p. 40S-47S.
57. Vongpatanasin, W., et al., *The Eisenmenger syndrome in adults.* Ann Intern Med, 1998. **128**(9): p. 745-55.
58. Engelfriet, P.M., et al., *Pulmonary arterial hypertension in adults born with a heart septal defect: the Euro Heart Survey on adult congenital heart disease.* Heart, 2007. **93**(6): p. 682-7.

59. Organization, W.H. *ICD-10: International statistical classification of diseases and related health problems*. 2008; Available from: sa9_thesis_corrected_3Jan2014.docx.
60. Leung, M.P., M.H. Tang, and A. Ghosh, *Prenatal diagnosis of congenital heart malformations: classification based on abnormalities detected by the four-chamber view.* Prenat Diagn, 1999. **19**(4): p. 305-13.
61. Knowles, R., et al., *Newborn screening for congenital heart defects: a systematic review and cost-effectiveness analysis.* Health Technol Assess, 2005. **9**(44): p. 1-152, iii-iv.
62. White, M.C., *Anaesthetic implications of congenital heart disease for children undergoing non-cardiac surgery.* Anaesthesia & Intensive Care Medicine, 2009. **10**(10): p. 504-509.
63. Connelly, M.S., et al., *Canadian Consensus Conference on Adult Congenital Heart Disease 1996.* Can J Cardiol, 1998. **14**(3): p. 395-452.
64. Lindinger, A., G. Schwedler, and H.W. Hense, *Prevalence of congenital heart defects in newborns in Germany: Results of the first registration year of the PAN Study (July 2006 to June 2007).* Klin Padiatr, 2010. **222**(5): p. 321-6.
65. Marino, B. and M.C. Digilio, *Congenital heart disease and genetic syndromes: specific correlation between cardiac phenotype and genotype.* Cardiovasc Pathol, 2000. **9**(6): p. 303-15.
66. Clark, E.B., *Pathogenetic mechanisms of congenital cardiovascular malformations revisited.* Semin Perinatol, 1996. **20**(6): p. 465-72.
67. (NCS), N.C.S. *OPCS-4 Classification*. 2011; Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/opcs4/.
68. Coding Committee of the Association for European Paediatric, C., *The European Paediatric Cardiac Code: the first revision.* Cardiol Young, 2002. **12 Suppl 2**: p. 1-211.
69. Vincent, S.D. and M.E. Buckingham, *How to make a heart: the origin and regulation of cardiac progenitor cells.* Curr Top Dev Biol, 2010. **90**: p. 1-41.
70. Epstein, J.A., *Franklin H. Epstein Lecture. Cardiac development and implications for heart disease.* N Engl J Med, 2010. **363**(17): p. 1638-47.
71. Yamagishi, H., et al., *Molecular embryology for an understanding of congenital heart diseases.* Anat Sci Int, 2009. **84**(3): p. 88-94.
72. Kelly, R.G., N.A. Brown, and M.E. Buckingham, *The arterial pole of the mouse heart forms from Fgf10-expressing cells in pharyngeal mesoderm.* Dev Cell, 2001. **1**(3): p. 435-40.
73. Buckingham, M., S. Meilhac, and S. Zaffran, *Building the mammalian heart from two sources of myocardial cells.* Nat Rev Genet, 2005. **6**(11): p. 826-35.
74. Srivastava, D. and E.N. Olson, *A genetic blueprint for cardiac development.* Nature, 2000. **407**(6801): p. 221-6.
75. Srivastava, D., *Making or breaking the heart: from lineage determination to morphogenesis.* Cell, 2006. **126**(6): p. 1037-48.
76. Hutson, M.R. and M.L. Kirby, *Model systems for the study of heart development and disease. Cardiac neural crest and conotruncal malformations.* Semin Cell Dev Biol, 2007. **18**(1): p. 101-10.

77. Waldo, K.L., et al., *Cardiac neural crest is necessary for normal addition of the myocardium to the arterial pole from the secondary heart field.* Dev Biol, 2005. **281**(1): p. 66-77.
78. Ward, C., et al., *Ablation of the secondary heart field leads to tetralogy of Fallot and pulmonary atresia.* Dev Biol, 2005. **284**(1): p. 72-83.
79. Mikawa, T. and R.G. Gourdie, *Pericardial mesoderm generates a population of coronary smooth muscle cells migrating into the heart along with ingrowth of the epicardial organ.* Dev Biol, 1996. **174**(2): p. 221-32.
80. Oostra, R.-J., G. Steding, and S. Virágh, *Steding's and Virágh's scanning electron microscopy atlas of the developing human heart*. 2007, New York: Springer. x, 211p.
81. Arraez-Aybar, L.A., A. Turrero-Nogues, and D.G. Marantos-Gamarra, *Embryonic cardiac morphometry in Carnegie stages 15-23, from the Complutense University of Madrid Institute of Embryology Human Embryo Collection.* Cells Tissues Organs, 2008. **187**(3): p. 211-20.
82. Sylva, M., M.J. van den Hoff, and A.F. Moorman, *Development of the Human Heart.* Am J Med Genet A, 2013: p. 0.
83. O'Rahilly, R., F. Müller, and G.L. Streeter, *Developmental stages in human embryos : including a revision of Streeter's "Horizons" and a survey of the Carnegie collection*. Publication / Carnegie Institution of Washington. 1987, Washington, D.C.: Carnegie Institution of Washington. 306 p., 1 leaf of plates.
84. Sommer, R.J., Z.M. Hijazi, and J.F. Rhodes, Jr., *Pathophysiology of congenital heart disease in the adult: part I: Shunt lesions.* Circulation, 2008. **117**(8): p. 1090-9.
85. Meissner, I., et al., *Patent foramen ovale: innocent or guilty? Evidence from a prospective population-based study.* J Am Coll Cardiol, 2006. **47**(2): p. 440-5.
86. Marie Valente, A. and J.F. Rhodes, *Current indications and contraindications for transcatheter atrial septal defect and patent foramen ovale device closure.* Am Heart J, 2007. **153**(4 Suppl): p. 81-4.
87. Heiden, K., *Congenital Heart Defects, Simplified* 2009: Midwest EchoSolutions.
88. Freed, M.D., et al., *Prostaglandin E1 infants with ductus arteriosus-dependent congenital heart disease.* Circulation, 1981. **64**(5): p. 899-905.
89. Audrey Marshall, M., *Hypoplastic left heart syndrome*. UpToDate.com, ed. D. Marion. 2013, Waltham, MA.
90. Jenkins, K.J., et al., *Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics.* Circulation, 2007. **115**(23): p. 2995-3014.
91. Wren, C., G. Birrell, and G. Hawthorne, *Cardiovascular malformations in infants of diabetic mothers.* Heart, 2003. **89**(10): p. 1217-20.
92. Cousins, L., *Etiology and prevention of congenital anomalies among infants of overt diabetic women.* Clin Obstet Gynecol, 1991. **34**(3): p. 481-93.
93. Levy, H.L., et al., *Congenital heart disease in maternal phenylketonuria: report from the Maternal PKU Collaborative Study.* Pediatr Res, 2001. **49**(5): p. 636-42.

94. Lenke, R.R. and H.L. Levy, *Maternal phenylketonuria and hyperphenylalaninemia. An international survey of the outcome of untreated and treated pregnancies.* N Engl J Med, 1980. **303**(21): p. 1202-8.
95. Botto, L.D., M.C. Lynberg, and J.D. Erickson, *Congenital heart defects, maternal febrile illness, and multivitamin use: a population-based study.* Epidemiology, 2001. **12**(5): p. 485-90.
96. Scanlon, K.S., et al., *Preconceptional folate intake and malformations of the cardiac outflow tract. Baltimore-Washington Infant Study Group.* Epidemiology, 1998. **9**(1): p. 95-8.
97. Stuckey, D., *Congenital heart defects following maternal rubella during pregnancy.* Br Heart J, 1956. **18**(4): p. 519-22.
98. Kelly, T.E., et al., *Teratogenicity of anticonvulsant drugs. II: A prospective study.* Am J Med Genet, 1984. **19**(3): p. 435-43.
99. Wilson, P.D., et al., *Attributable fraction for cardiac malformations.* Am J Epidemiol, 1998. **148**(5): p. 414-23.
100. Geiger, J.M., M. Baudin, and J.H. Saurat, *Teratogenic risk with etretinate and acitretin treatment.* Dermatology, 1994. **189**(2): p. 109-16.
101. Pierpont, M.E., et al., *Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics.* Circulation, 2007. **115**(23): p. 3015-38.
102. Roos-Hesselink JW, K.-F.W., Meijboom FJ, Pieper PG. , *Inheritance of congenital heart disease.* Neth Heart J 2005. **13: 88-91**.
103. Merscher, S., et al., *TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome.* Cell, 2001. **104**(4): p. 619-29.
104. Lange, A.W., J.D. Molkentin, and K.E. Yutzey, *DSCR1 gene expression is dependent on NFATc1 during cardiac valve formation and colocalizes with anomalous organ development in trisomy 16 mice.* Dev Biol, 2004. **266**(2): p. 346-60.
105. Arron, J.R., et al., *NFAT dysregulation by increased dosage of DSCR1 and DYRK1A on chromosome 21.* Nature, 2006. **441**(7093): p. 595-600.
106. Subramaniam, P., et al., *Diagnosis of Alagille syndrome-25 years of experience at King's College Hospital.* J Pediatr Gastroenterol Nutr, 2011. **52**(1): p. 84-9.
107. Warthen, D.M., et al., *Jagged1 (JAG1) mutations in Alagille syndrome: increasing the mutation detection rate.* Hum Mutat, 2006. **27**(5): p. 436-43.
108. Krantz, I.D., et al., *Deletions of 20p12 in Alagille syndrome: frequency and molecular characterization.* Am J Med Genet, 1997. **70**(1): p. 80-6.
109. Schott, J.J., et al., *Congenital heart disease caused by mutations in the transcription factor NKX2-5.* Science, 1998. **281**(5373): p. 108-11.
110. Garg, V., et al., *GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5.* Nature, 2003. **424**(6947): p. 443-7.
111. Wessels, M.W. and P.J. Willems, *Genetic factors in non-syndromic congenital heart malformations.* Clin Genet, 2010. **78**(2): p. 103-23.

112. Cordell, H.J., et al., *Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16.* Nat Genet, 2013. **45**(7): p. 822-4.
113. Goodship, J.A., et al., *A common variant in the PTPN11 gene contributes to the risk of tetralogy of Fallot.* Circ Cardiovasc Genet, 2012. **5**(3): p. 287-92.
114. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
115. Hu, Z., et al., *A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations.* Nat Genet, 2013. **45**(7): p. 818-21.
116. Reamon-Buettner, S.M. and J. Borlak, *Somatic NKX2-5 mutations as a novel mechanism of disease in complex congenital heart disease.* J Med Genet, 2004. **41**(9): p. 684-90.
117. Reamon-Buettner, S.M. and J. Borlak, *TBX5 mutations in non-Holt-Oram syndrome (HOS) malformed hearts.* Hum Mutat, 2004. **24**(1): p. 104.
118. Draus, J.M., Jr., et al., *Investigation of somatic NKX2-5 mutations in congenital heart disease.* J Med Genet, 2009. **46**(2): p. 115-22.
119. Cordes, K.R. and D. Srivastava, *MicroRNA regulation of cardiovascular development.* Circ Res, 2009. **104**(6): p. 724-32.
120. Liu, N. and E.N. Olson, *MicroRNA regulatory networks in cardiovascular development.* Dev Cell, 2010. **18**(4): p. 510-25.
121. Zhu, S., et al., *Identification of maternal serum microRNAs as novel non-invasive biomarkers for prenatal detection of fetal congenital heart defects.* Clin Chim Acta, 2013. **424C**: p. 66-72.
122. Soemedi, R., et al., *Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease.* Am J Hum Genet, 2012. **91**(3): p. 489-501.
123. Zaidi, S., et al., *De novo mutations in histone-modifying genes in congenital heart disease.* Nature, 2013. **498**(7453): p. 220-3.
124. Bentham, J. and S. Bhattacharya, *Genetic mechanisms controlling cardiovascular development.* Ann N Y Acad Sci, 2008. **1123**: p. 10-9.
125. Hutchison, C.A., 3rd, *DNA sequencing: bench to bedside and beyond.* Nucleic Acids Res, 2007. **35**(18): p. 6227-37.
126. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
127. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
128. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-32.
129. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
130. Blazej, R.G., P. Kumaresan, and R.A. Mathies, *Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing.* Proc Natl Acad Sci U S A, 2006. **103**(19): p. 7240-5.
131. Gresham, D., M.J. Dunham, and D. Botstein, *Comparing whole genomes using DNA microarrays.* Nat Rev Genet, 2008. **9**(4): p. 291-302.
132. Healy, K., *Nanopore-based single-molecule DNA analysis.* Nanomedicine (Lond), 2007. **2**(4): p. 459-81.

133. Soni, G.V. and A. Meller, *Progress toward ultrafast DNA sequencing using solid-state nanopores.* Clin Chem, 2007. **53**(11): p. 1996-2001.
134. Mitra, R.D. and G.M. Church, *In situ localized amplification and contact replication of many individual DNA molecules.* Nucleic Acids Res, 1999. **27**(24): p. e34.
135. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
136. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
137. Pabinger, S., et al., *A survey of tools for variant analysis of next-generation genome sequencing data.* Brief Bioinform, 2013.
138. Knierim, E., et al., *Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing.* PLoS One, 2011. **6**(11): p. e28240.
139. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
140. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.
141. Glenn, T. *2013 NGS Field Guide*. 2013; Available from: http://www.molecularecologist.com/next-gen-fieldguide-2013/.
142. Cock, P.J., et al., *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.* Nucleic Acids Res, 2010. **38**(6): p. 1767-71.
143. Nielsen, R., et al., *Genotype and SNP calling from next-generation sequencing data.* Nat Rev Genet, 2011. **12**(6): p. 443-51.
144. Dai, M., et al., *NGSQC: cross-platform quality analysis pipeline for deep sequencing data.* BMC Genomics, 2010. **11 Suppl 4**: p. S7.
145. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Brief Bioinform, 2010. **11**(5): p. 473-83.
146. Ruffalo, M., T. LaFramboise, and M. Koyuturk, *Comparative analysis of algorithms for next-generation sequencing read alignment.* Bioinformatics, 2011. **27**(20): p. 2790-6.
147. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
148. Ning, Z., A.J. Cox, and J.C. Mullikin, *SSAHA: a fast search method for large DNA databases.* Genome Res, 2001. **11**(10): p. 1725-9.
149. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
150. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
151. Malhis, N., et al., *Slider--maximum use of probability information for alignment of short sequence reads and SNP detection.* Bioinformatics, 2009. **25**(1): p. 6-13.
152. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
153. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

154. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
155. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
156. Kim, S.Y. and T.P. Speed, *Comparing somatic mutation-callers: beyond Venn diagrams.* BMC Bioinformatics, 2013. **14**: p. 189.
157. Duan, J., et al., *Comparative studies of copy number variation detection methods for next-generation sequencing technologies.* PLoS One, 2013. **8**(3): p. e59128.
158. Albers, C.A., et al., *Dindel: accurate indel calls from short-read data.* Genome Res, 2011. **21**(6): p. 961-73.
159. Ye, K., et al., *Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.* Bioinformatics, 2009. **25**(21): p. 2865-71.
160. Neuman, J.A., O. Isakov, and N. Shomron, *Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection.* Brief Bioinform, 2013. **14**(1): p. 46-55.
161. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
162. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files.* Bioinformatics, 2011. **27**(5): p. 718-9.
163. Tennessen, J.A., et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes.* Science, 2012. **337**(6090): p. 64-9.
164. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.
165. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.
166. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
167. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.
168. Stenson, P.D., et al., *The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 13.
169. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
170. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-70.
171. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
172. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

173. Gonzalez-Perez, A. and N. Lopez-Bigas, *Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.* Am J Hum Genet, 2011. **88**(4): p. 440-9.
174. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.
175. Dewey, F.E., et al., *DNA sequencing: clinical applications of new DNA sequencing technologies.* Circulation, 2012. **125**(7): p. 931-44.
176. Johnson, J.O., et al., *Exome sequencing reveals VCP mutations as a cause of familial ALS.* Neuron, 2010. **68**(5): p. 857-64.
177. Bonnefond, A., et al., *Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome.* PLoS One, 2010. **5**(10): p. e13630.
178. Ostergaard, P., et al., *Rapid identification of mutations in GJC2 in primary lymphoedema using whole exome sequencing combined with linkage analysis with delineation of the phenotype.* J Med Genet, 2011. **48**(4): p. 251-5.
179. Wang, J.L., et al., *TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing.* Brain, 2010. **133**(Pt 12): p. 3510-8.
180. Sirmaci, A., et al., *MASP1 mutations in patients with facial, umbilical, coccygeal, and auditory findings of Carnevale, Malpuech, OSA, and Michels syndromes.* Am J Hum Genet, 2010. **87**(5): p. 679-86.
181. Montenegro, G., et al., *Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family.* Ann Neurol, 2011. **69**(3): p. 464-70.
182. Choi, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19096-101.
183. Bolze, A., et al., *Whole-exome-sequencing-based discovery of human FADD deficiency.* Am J Hum Genet, 2010. **87**(6): p. 873-81.
184. Musunuru, K., et al., *Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia.* N Engl J Med, 2010. **363**(23): p. 2220-7.
185. Lalonde, E., et al., *Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing.* Hum Mutat, 2010. **31**(8): p. 918-23.
186. Edvardson, S., et al., *Joubert syndrome 2 (JBTS2) in Ashkenazi Jews is associated with a TMEM216 mutation.* Am J Hum Genet, 2010. **86**(1): p. 93-7.
187. Caliskan, M., et al., *Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13.* Hum Mol Genet, 2011. **20**(7): p. 1285-9.
188. Walsh, T., et al., *Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82.* Am J Hum Genet, 2010. **87**(1): p. 90-4.
189. Kalay, E., et al., *CEP152 is a genome maintenance protein disrupted in Seckel syndrome.* Nat Genet, 2011. **43**(1): p. 23-6.
190. Vissers, L.E., et al., *A de novo paradigm for mental retardation.* Nat Genet, 2010. **42**(12): p. 1109-12.

191. Hoischen, A., et al., *De novo mutations of SETBP1 cause Schinzel-Giedion syndrome.* Nat Genet, 2010. **42**(6): p. 483-5.
192. Worthey, E.A., et al., *Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease.* Genet Med, 2011. **13**(3): p. 255-62.
193. Sobreira, N.L., et al., *Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene.* PLoS Genet, 2010. **6**(6): p. e1000991.
194. Lupski, J.R., et al., *Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.* N Engl J Med, 2010. **362**(13): p. 1181-91.
195. Roach, J.C., et al., *Analysis of genetic inheritance in a family quartet by whole-genome sequencing.* Science, 2010. **328**(5978): p. 636-9.
196. Rios, J., et al., *Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia.* Hum Mol Genet, 2010. **19**(22): p. 4313-8.
197. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
198. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
199. Fu, W., et al., *Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.* Nature, 2013. **493**(7431): p. 216-20.
200. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
201. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
202. Boycott, K.M., et al., *Rare-disease genetics in the era of next-generation sequencing: discovery to translation.* Nat Rev Genet, 2013.
203. Schuurs-Hoeijmakers, J.H., et al., *Mutations in DDHD2, encoding an intracellular phospholipase A(1), cause a recessive form of complex hereditary spastic paraplegia.* Am J Hum Genet, 2012. **91**(6): p. 1073-81.
204. Kalsoom, U.E., et al., *Whole exome sequencing identified a novel zinc-finger gene ZNF141 associated with autosomal recessive postaxial polydactyly type A.* J Med Genet, 2013. **50**(1): p. 47-53.
205. Sankaran, V.G., et al., *Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia.* J Clin Invest, 2012. **122**(7): p. 2439-43.
206. Fiskerstrand, T., et al., *Familial diarrhea syndrome caused by an activating GUCY2C mutation.* N Engl J Med, 2012. **366**(17): p. 1586-95.
207. Gibson, W.T., et al., *Mutations in EZH2 cause Weaver syndrome.* Am J Hum Genet, 2012. **90**(1): p. 110-8.
208. Boyd, S.D., *Diagnostic applications of high-throughput DNA sequencing.* Annu Rev Pathol, 2013. **8**: p. 381-410.
209. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.
210. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet, 2008. **40**(6): p. 722-9.
211. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-90.

212. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.
213. Ley, T.J., et al., *DNMT3A mutations in acute myeloid leukemia.* N Engl J Med, 2010. **363**(25): p. 2424-33.
214. Logan, A.C., et al., *High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment.* Proc Natl Acad Sci U S A, 2011. **108**(52): p. 21194-9.
215. Boyd, S.D., et al., *Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing.* Sci Transl Med, 2009. **1**(12): p. 12ra23.
216. Cohorts for, H., et al., *Whole-genome sequence-based analysis of high-density lipoprotein cholesterol.* Nat Genet, 2013. **45**(8): p. 899-901.
217. Chin, C.S., et al., *The origin of the Haitian cholera outbreak strain.* N Engl J Med, 2011. **364**(1): p. 33-42.
218. Rasko, D.A., et al., *Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany.* N Engl J Med, 2011. **365**(8): p. 709-17.
219. Assiri, A., et al., *Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus.* N Engl J Med, 2013.
220. Snyder, T.M., et al., *Universal noninvasive detection of solid organ transplant rejection.* Proc Natl Acad Sci U S A, 2011. **108**(15): p. 6229-34.
221. Fan, H.C., et al., *Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood.* Proc Natl Acad Sci U S A, 2008. **105**(42): p. 16266-71.
222. Chiu, R.W., et al., *Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study.* BMJ, 2011. **342**: p. c7401.
223. Bornman, D.M., et al., *Short-read, high-throughput sequencing technology for STR genotyping.* Biotechniques, 2012. **0**(0): p. 1-6.
224. Warshauer, D.H., et al., *STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data.* Forensic Sci Int Genet, 2013. **7**(4): p. 409-17.
225. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
226. Soon, W.W., M. Hariharan, and M.P. Snyder, *High-throughput sequencing for biology and medicine.* Mol Syst Biol, 2013. **9**: p. 640.
227. Waern, K., U. Nagalakshmi, and M. Snyder, *RNA sequencing.* Methods Mol Biol, 2011. **759**: p. 125-32.
228. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nat Methods, 2006. **3**(3): p. 211-22.
229. Fullwood, M.J., et al., *Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.* Genome Res, 2009. **19**(4): p. 521-32.
230. Chu, C., et al., *Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.* Mol Cell, 2011. **44**(4): p. 667-78.
231. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.

232. Churchman, L.S. and J.S. Weissman, *Nascent transcript sequencing visualizes transcription at nucleotide resolution.* Nature, 2011. **469**(7330): p. 368-73.
233. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.
234. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
235. Hesselberth, J.R., et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods, 2009. **6**(4): p. 283-9.
236. Crawford, G.E., et al., *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Res, 2006. **16**(1): p. 123-31.
237. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.* Genome Res, 2007. **17**(6): p. 877-85.
238. Wang, Z., et al., *Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes.* Cell, 2009. **138**(5): p. 1019-31.
239. Smith, Z.D., et al., *High-throughput bisulfite sequencing in mammalian genomes.* Methods, 2009. **48**(3): p. 226-32.
240. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.* Genome Res, 2006. **16**(10): p. 1299-309.
241. Fullwood, M.J., et al., *An oestrogen-receptor-alpha-bound human chromatin interactome.* Nature, 2009. **462**(7269): p. 58-64.
242. Stein, L.D., *The case for cloud computing in genome informatics.* Genome Biol, 2010. **11**(5): p. 207.
243. Lin, Z., A.B. Owen, and R.B. Altman, *Genetics. Genomic research and human subject privacy.* Science, 2004. **305**(5681): p. 183.
244. Homer, N., et al., *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.* PLoS Genet, 2008. **4**(8): p. e1000167.
245. Greenbaum, D., et al., *Genomics and privacy: implications of the new reality of closed data for the field.* PLoS Comput Biol, 2011. **7**(12): p. e1002278.
246. Rehm, H.L., *Disease-targeted sequencing: a cornerstone in the clinic.* Nat Rev Genet, 2013. **14**(4): p. 295-300.
247. Green, R.C., et al., *Exploring concordance and discordance for return of incidental findings from clinical sequencing.* Genet Med, 2012. **14**(4): p. 405-10.
248. Makrythanasis, P. and S.E. Antonarakis, *High-throughput sequencing and rare genetic diseases.* Mol Syndromol, 2012. **3**(5): p. 197-203.
249. Stevenson, D.A. and J.C. Carey, *Contribution of malformations and genetic disorders to mortality in a children's hospital.* Am J Med Genet A, 2004. **126A**(4): p. 393-7.
250. Yoon, P.W., et al., *Contribution of birth defects and genetic diseases to pediatric hospitalizations. A population-based study.* Arch Pediatr Adolesc Med, 1997. **151**(11): p. 1096-103.

251. Kumar, P., et al., *Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department.* Mayo Clin Proc, 2001. **76**(8): p. 777-83.
252. Barroso, I., et al., *Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension.* Nature, 1999. **402**(6764): p. 880-3.
253. Amberger, J., et al., *McKusick's Online Mendelian Inheritance in Man (OMIM).* Nucleic Acids Res, 2009. **37**(Database issue): p. D793-6.
254. Rabbani, B., et al., *Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders.* J Hum Genet, 2012. **57**(10): p. 621-32.
255. Kaasinen, E., et al., *Recessively inherited right atrial isomerism caused by mutations in growth/differentiation factor 1 (GDF1).* Hum Mol Genet, 2010. **19**(14): p. 2747-53.
256. Zaidi, S., et al., *De novo mutations in histone-modifying genes in congenital heart disease.* Nature, 2013.
257. Cordell, H.J., et al., *Genome-wide association study identifies loci on 12q24 and 13q32 associated with Tetralogy of Fallot.* Hum Mol Genet, 2013.
258. Olander, E., et al., *Third Prader-Willi syndrome phenotype due to maternal uniparental disomy 15 with mosaic trisomy 15.* Am J Med Genet, 2000. **93**(3): p. 215-8.
259. Wang, W., et al., *MTHFR C677T polymorphism and risk of congenital heart defects: evidence from 29 case-control and TDT studies.* PLoS One, 2013. **8**(3): p. e58041.
260. Firth, H.V., C.F. Wright, and D.D.D. Study, *The Deciphering Developmental Disorders (DDD) study.* Dev Med Child Neurol, 2011. **53**(8): p. 702-3.
261. Ge, D., et al., *SVA: software for annotating and visualizing sequenced human genomes.* Bioinformatics, 2011. **27**(14): p. 1998-2000.
262. Coutant, S., et al., *EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics.* BMC Bioinformatics, 2012. **13 Suppl 14**: p. S9.
263. Teer, J.K., et al., *VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer.* Bioinformatics, 2012. **28**(4): p. 599-600.
264. UK10K. *UK10K*. 2013; Available from: http://www.uk10k.org.
265. Conrad, D.F., et al., *Variation in genome-wide mutation rates within and between human families.* Nat Genet, 2011. **43**(7): p. 712-4.
266. Ramu, A., et al., *DeNovoGear: de novo indel and point mutation discovery and phasing.* Nat Methods, 2013.
267. Sanders, S.J., et al., *De novo mutations revealed by whole-exome sequencing are strongly associated with autism.* Nature, 2012. **485**(7397): p. 237-41.
268. O'Roak, B.J., et al., *Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations.* Nature, 2012. **485**(7397): p. 246-50.
269. Iossifov, I., et al., *De novo gene disruptions in children on the autistic spectrum.* Neuron, 2012. **74**(2): p. 285-99.
270. Neale, B.M., et al., *Patterns and rates of exonic de novo mutations in autism spectrum disorders.* Nature, 2012. **485**(7397): p. 242-5.

271. Rauch, A., et al., *Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study.* Lancet, 2012. **380**(9854): p. 1674-82.
272. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
273. Huang, N., et al., *Characterising and predicting haploinsufficiency in the human genome.* PLoS Genet, 2010. **6**(10): p. e1001154.
274. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
275. *GATK Technical Documentation*. 2013 GATK version 2.7-2-g701cd16 built at 2013/08/28 16:38:05.; Available from: http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_QualByDepth.html.
276. Albers, C.A., et al., *Dindel: Accurate indel calls from short-read data.* Genome Res, 2010.
277. Ahn, S.J., J. Costa, and J.R. Emanuel, *PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR.* Nucleic Acids Res, 1996. **24**(13): p. 2623-5.
278. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
279. Dick, K.J., et al., *Refinement of the locus for distal hereditary motor neuronopathy VII (dHMN-VII) and exclusion of candidate genes.* Genome, 2008. **51**(11): p. 959-62.
280. McEntagart, M., et al., *Localization of the gene for distal hereditary motor neuronopathy VII (dHMN-VII) to chromosome 2q14.* Am J Hum Genet, 2001. **68**(5): p. 1270-6.
281. Barwick, K.E., et al., *Defective presynaptic choline transport underlies hereditary motor neuropathy.* Am J Hum Genet, 2012. **91**(6): p. 1103-7.
282. Baple, E.L., et al., *Mutations in KPTN Cause Macrocephaly, Neurodevelopmental Delay, and Seizures.* Am J Hum Genet, 2013.
283. Harlalka, G.V., et al., *Mutations in B4GALNT1 (GM2 synthase) underlie a new disorder of ganglioside biosynthesis.* Brain, 2013. **136**(Pt 12): p. 3618-24.
284. Asai, K., et al., *Isolation of novel human cDNA (hGMF-gamma) homologous to Glia Maturation Factor-beta gene.* Biochim Biophys Acta, 1998. **1396**(3): p. 242-4.
285. Ikeda, K., et al., *Glia maturation factor-gamma is preferentially expressed in microvascular endothelial and inflammatory cells and modulates actin cytoskeleton reorganization.* Circ Res, 2006. **99**(4): p. 424-33.
286. Walker, M.G., *Gene expression versus sequence for predicting function: Glia Maturation Factor gamma is not a glia maturation factor.* Genomics Proteomics Bioinformatics, 2003. **1**(1): p. 52-7.
287. Sleep, E., et al., *Transcriptomics approach to investigate zebrafish heart regeneration.* J Cardiovasc Med (Hagerstown), 2010. **11**(5): p. 369-80.
288. Eppig, J.T., et al., *The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D881-6.

289. Merveille, A.C., et al., *CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs.* Nat Genet, 2011. **43**(1): p. 72-8.
290. Olbrich, H., et al., *Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry.* Am J Hum Genet, 2012. **91**(4): p. 672-84.
291. McInerney-Leo, A.M., et al., *Short-Rib Polydactyly and Jeune Syndromes Are Caused by Mutations in WDR60.* Am J Hum Genet, 2013.
292. Schmidts, M., et al., *Combined NGS approaches identify mutations in the intraflagellar transport gene IFT140 in skeletal ciliopathies with early progressive kidney Disease.* Hum Mutat, 2013. **34**(5): p. 714-24.
293. Fallot., E.L.A., *Contribution i l'anatomie pathologique de la maladie bleue* Marseille médical, 1888. **25: 77-93, 138-158, 207-223, 341-354, 370-386, 403-420**.
294. M. Cristina Digilio, B.D., Bruno Marino, *The right ventricle in adults with tetralogy of fallot*, ed. A.G. Massimo Chessa. 2012, New York: Springer.
295. Abbott ME, D.W., *The clinical classification of congenital heart disease, with remarks upon its pathological anatomy, diagnosis and treatment.* Int Clin, 1924. **4:156-188**.
296. Abbott, M.E., *Atlas of congenital cardiac disease*. 1936, New York, N.Y.,: The American heart association. x, 62 p. incl. front. (5 port.) illus., diagrs.
297. Ferencz, C., Rubin, JD, Loffredo,CA, Magee,CM., *The Epidemiology of Congenital Heart Disease, The Baltimore-Washington Infant Study (1981-1989)*. Perspectives in Pediatric Cardiology. Vol. vol.4. . 1993: Futura Publishing Co.Inc.
298. Anderson RH, M.F., Shinebourne EA, *Fallot's Tetralogy. In: Paediatric Cardiology*, ed. T. M. 2002: London: Churchill Livingstone.
299. Jenkins, K.J., et al., *Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics.* Circulation, 2007. **115**(23): p. 2995-3014.
300. Correa-Villasenor, A., et al., *White-black differences in cardiovascular malformations in infancy and socioeconomic factors. The Baltimore-Washington Infant Study Group.* Am J Epidemiol, 1991. **134**(4): p. 393-402.
301. Digilio, M.C., et al., *Recurrence risk figures for isolated tetralogy of Fallot after screening for 22q11 microdeletion.* J Med Genet, 1997. **34**(3): p. 188-90.
302. Bailliard, F. and R.H. Anderson, *Tetralogy of Fallot.* Orphanet J Rare Dis, 2009. **4**: p. 2.
303. Hansen, J.T. and F.H. Netter, *Netter's clinical anatomy*. 2nd ed. 2010, Philadelphia: Saunders/Elsevier. xviii, 470 p.
304. Anderson, R.H. and P.M. Weinberg, *The clinical anatomy of tetralogy of fallot.* Cardiol Young, 2005. **15 Suppl 1**: p. 38-47.
305. Jiang, X., et al., *Fate of the mammalian cardiac neural crest.* Development, 2000. **127**(8): p. 1607-16.

306. Anderson, R.H., et al., *Development of the heart: (3) formation of the ventricular outflow tracts, arterial valves, and intrapericardial arterial trunks.* Heart, 2003. **89**(9): p. 1110-8.
307. Lin, C.J., et al., *Partitioning the heart: mechanisms of cardiac septation and valve development.* Development, 2012. **139**(18): p. 3277-99.
308. Changela, V., C. John, and S. Maheshwari, *Unusual cardiac associations with Tetralogy of Fallot-a descriptive study.* Pediatr Cardiol, 2010. **31**(6): p. 785-91.
309. Gilboa, S.M., et al., *Relation between ambient air quality and selected birth defects, seven county study, Texas, 1997-2000.* Am J Epidemiol, 2005. **162**(3): p. 238-52.
310. Digilio, M.C., et al., *Comparison of occurrence of genetic syndromes in ventricular septal defect with pulmonic stenosis (classic tetralogy of Fallot) versus ventricular septal defect with pulmonic atresia.* Am J Cardiol, 1996. **77**(15): p. 1375-6.
311. Freeman, S.B., et al., *Ethnicity, sex, and the incidence of congenital heart defects: a report from the National Down Syndrome Project.* Genet Med, 2008. **10**(3): p. 173-80.
312. Karr, S.S., et al., *Tetralogy of Fallot. The spectrum of severity in a regional study, 1981-1985.* Am J Dis Child, 1992. **146**(1): p. 121-4.
313. Musewe, N.N., et al., *Echocardiographic evaluation of the spectrum of cardiac anomalies associated with trisomy 13 and trisomy 18.* J Am Coll Cardiol, 1990. **15**(3): p. 673-7.
314. McDonald-McGinn, D.M., et al., *The Philadelphia story: the 22q11.2 deletion: report on 250 patients.* Genet Couns, 1999. **10**(1): p. 11-24.
315. Ryan, A.K., et al., *Spectrum of clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study.* J Med Genet, 1997. **34**(10): p. 798-804.
316. Lindsay, E.A., et al., *Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice.* Nature, 2001. **410**(6824): p. 97-101.
317. McElhinney, D.B., et al., *Analysis of cardiovascular phenotype and genotype-phenotype correlation in individuals with a JAG1 mutation and/or Alagille syndrome.* Circulation, 2002. **106**(20): p. 2567-74.
318. Emerick, K.M., et al., *Features of Alagille syndrome in 92 patients: frequency and relation to prognosis.* Hepatology, 1999. **29**(3): p. 822-9.
319. Spinner, N.B., L.D. Leonard, and I.D. Krantz, *Alagille Syndrome*, in *GeneReviews*, R.A. Pagon, et al., Editors. 1993: Seattle (WA).
320. Crosnier, C., et al., *Mutations in JAGGED1 gene are predominantly sporadic in Alagille syndrome.* Gastroenterology, 1999. **116**(5): p. 1141-8.
321. Bauer, R.C., et al., *Jagged1 (JAG1) mutations in patients with tetralogy of Fallot or pulmonic stenosis.* Hum Mutat, 2010. **31**(5): p. 594-601.
322. Rauch, R., et al., *Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot.* J Med Genet, 2010. **47**(5): p. 321-31.
323. Eldadah, Z.A., et al., *Familial Tetralogy of Fallot caused by mutation in the jagged1 gene.* Hum Mol Genet, 2001. **10**(2): p. 163-9.
324. Krantz, I.D., et al., *Spectrum and frequency of jagged1 (JAG1) mutations in Alagille syndrome patients and their families.* Am J Hum Genet, 1998. **62**(6): p. 1361-9.

325. Lu, F., J.J. Morrissette, and N.B. Spinner, *Conditional JAG1 mutation shows the developing heart is more sensitive than developing liver to JAG1 dosage.* Am J Hum Genet, 2003. **72**(4): p. 1065-70.
326. Majewski, J., et al., *Mutations in NOTCH2 in families with Hajdu-Cheney syndrome.* Hum Mutat, 2011. **32**(10): p. 1114-7.
327. Zanotti, S. and E. Canalis, *Notch and the skeleton.* Mol Cell Biol, 2010. **30**(4): p. 886-96.
328. Penton, A.L., L.D. Leonard, and N.B. Spinner, *Notch signaling in human development and disease.* Semin Cell Dev Biol, 2012. **23**(4): p. 450-7.
329. Blake, K.D. and C. Prasad, *CHARGE syndrome.* Orphanet J Rare Dis, 2006. **1**: p. 34.
330. Jay, P.Y., et al., *Nkx2-5 mutation causes anatomic hypoplasia of the cardiac conduction system.* J Clin Invest, 2004. **113**(8): p. 1130-7.
331. McElhinney, D.B., et al., *NKX2.5 mutations in patients with congenital heart disease.* J Am Coll Cardiol, 2003. **42**(9): p. 1650-5.
332. Goldmuntz, E., E. Geiger, and D.W. Benson, *NKX2.5 mutations in patients with tetralogy of fallot.* Circulation, 2001. **104**(21): p. 2565-8.
333. Pizzuti, A., et al., *Mutations of ZFPM2/FOG2 gene in sporadic cases of tetralogy of Fallot.* Hum Mutat, 2003. **22**(5): p. 372-7.
334. Sperling, S., et al., *Identification and functional analysis of CITED2 mutations in patients with congenital heart defects.* Hum Mutat, 2005. **26**(6): p. 575-82.
335. Roessler, E., et al., *Reduced NODAL signaling strength via mutation of several pathway members including FOXH1 is linked to human heart defects and holoprosencephaly.* Am J Hum Genet, 2008. **83**(1): p. 18-29.
336. Guida, V., et al., *Novel and recurrent JAG1 mutations in patients with tetralogy of Fallot.* Clin Genet, 2011. **80**(6): p. 591-4.
337. Griffin, H.R., et al., *Systematic survey of variants in TBX1 in non-syndromic tetralogy of Fallot identifies a novel 57 base pair deletion that reduces transcriptional activity but finds no evidence for association with common variants.* Heart, 2010. **96**(20): p. 1651-5.
338. A Töpf, H.R.G., D H Hall, E Glen, B D Keavney, J A Goodship, The Change Study Collaborators, *Gene screening of the secondary heart field network in tetralogy of fallot patients.* Heart, 2011.
339. Guida, V., et al., *A variant in the carboxyl-terminus of connexin 40 alters GAP junctions and increases risk for tetralogy of Fallot.* Eur J Hum Genet, 2013. **21**(1): p. 69-75.
340. Silversides, C.K., et al., *Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways.* PLoS Genet, 2012. **8**(8): p. e1002843.
341. Greenway, S.C., et al., *De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot.* Nat Genet, 2009. **41**(8): p. 931-5.
342. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* Curr Opin Genet Dev, 2009. **19**(3): p. 212-9.
343. Pritchard, J.K. and N.J. Cox, *The allelic architecture of human disease genes: common disease-common variant...or not?* Hum Mol Genet, 2002. **11**(20): p. 2417-23.

344. Cordell, H.J., et al., *Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot.* Hum Mol Genet, 2013. **22**(7): p. 1473-81.
345. Smyth, D.J., et al., *Shared and distinct genetic variants in type 1 diabetes and celiac disease.* N Engl J Med, 2008. **359**(26): p. 2767-77.
346. Soranzo, N., et al., *A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium.* Nat Genet, 2009. **41**(11): p. 1182-90.
347. Stahl, E.A., et al., *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci.* Nat Genet, 2010. **42**(6): p. 508-14.
348. Tartaglia, M., et al., *Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome.* Nat Genet, 2001. **29**(4): p. 465-8.
349. Filmus, J., M. Capurro, and J. Rast, *Glypicans.* Genome Biol, 2008. **9**(5): p. 224.
350. E., D.-G., *Hypothèses de dimérie et de non-pénétrance.* Acta genet, 1962. **12**: p. 65–96
.
351. Schaffer, A.A., *Digenic inheritance in medical genetics.* J Med Genet, 2013.
352. Kajiwara, K., E.L. Berson, and T.P. Dryja, *Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci.* Science, 1994. **264**(5165): p. 1604-8.
353. Lemmers, R.J., et al., *Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2.* Nat Genet, 2012. **44**(12): p. 1370-4.
354. Margolin, D.H., et al., *Ataxia, dementia, and hypogonadotropism caused by disordered ubiquitination.* N Engl J Med, 2013. **368**(21): p. 1992-2003.
355. You, F.M., et al., *BatchPrimer3: a high throughput web application for PCR and sequencing primer design.* BMC Bioinformatics, 2008. **9**: p. 253.
356. *Geneious* Biomatters.
357. Kryukov, G.V., L.A. Pennacchio, and S.R. Sunyaev, *Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.* Am J Hum Genet, 2007. **80**(4): p. 727-39.
358. Bray, S.J., *Notch signalling: a simple pathway becomes complex.* Nat Rev Mol Cell Biol, 2006. **7**(9): p. 678-89.
359. McBride, K.L., et al., *NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling.* Hum Mol Genet, 2008. **17**(18): p. 2886-93.
360. Mohamed, S.A., et al., *Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve.* Biochem Biophys Res Commun, 2006. **345**(4): p. 1460-5.
361. Garg, V., et al., *Mutations in NOTCH1 cause aortic valve disease.* Nature, 2005. **437**(7056): p. 270-4.
362. Mao, Y., et al., *Characterization of a Dchs1 mutant mouse reveals requirements for Dchs1-Fat4 signaling during mammalian development.* Development, 2011. **138**(5): p. 947-57.
363. Kuroda, K., et al., *Regulation of marginal zone B cell development by MINT, a suppressor of Notch/RBP-J signaling pathway.* Immunity, 2003. **18**(2): p. 301-12.

364. Gocke, C.B. and H. Yu, *ZNF198 stabilizes the LSD1-CoREST-HDAC1 complex on chromatin through its MYM-type zinc fingers.* PLoS One, 2008. **3**(9): p. e3255.
365. Xiao, S., et al., *FGFR1 is fused with a novel zinc-finger gene, ZNF198, in the t(8;13) leukaemia/lymphoma syndrome.* Nat Genet, 1998. **18**(1): p. 84-7.
366. Ren, M. and J.K. Cowell, *Constitutive Notch pathway activation in murine ZMYM2-FGFR1-induced T-cell lymphomas associated with atypical myeloproliferative disease.* Blood, 2011. **117**(25): p. 6837-47.
367. Puck, J.M. and H.F. Willard, *X inactivation in females with X-linked disease.* N Engl J Med, 1998. **338**(5): p. 325-8.
368. Kawagoe, T., et al., *Sequential control of Toll-like receptor-dependent responses by IRAK1 and IRAK2.* Nat Immunol, 2008. **9**(6): p. 684-91.
369. Ramalingam, T.R., et al., *Unique functions of the type II interleukin 4 receptor identified in mice lacking the interleukin 13 receptor alpha1 chain.* Nat Immunol, 2008. **9**(1): p. 25-33.
370. Christensen, S.R., et al., *Toll-like receptor 7 and TLR9 dictate autoantibody specificity and have opposing inflammatory and regulatory roles in a murine model of lupus.* Immunity, 2006. **25**(3): p. 417-28.
371. Lugtenberg, D., et al., *ZNF674: a new kruppel-associated box-containing zinc-finger gene involved in nonsyndromic X-linked mental retardation.* Am J Hum Genet, 2006. **78**(2): p. 265-78.
372. Hurles, P.V.a.M. *CoNVex*. 2013; Available from: /nfs/users/nfs_p/pv1/ConvexPackage/CoNVex_0.5.tar.gz.
373. Grozinger, C.M., C.A. Hassig, and S.L. Schreiber, *Three proteins define a class of human histone deacetylases related to yeast Hda1p.* Proc Natl Acad Sci U S A, 1999. **96**(9): p. 4868-73.
374. Vega, R.B., et al., *Histone deacetylase 4 controls chondrocyte hypertrophy during skeletogenesis.* Cell, 2004. **119**(4): p. 555-66.
375. Aldred, M.A., et al., *Molecular analysis of 20 patients with 2q37.3 monosomy: definition of minimum deletion intervals for key phenotypes.* J Med Genet, 2004. **41**(6): p. 433-9.
376. Williams, S.R., et al., *Haploinsufficiency of HDAC4 causes brachydactyly mental retardation syndrome, with brachydactyly type E, developmental delays, and behavioral problems.* Am J Hum Genet, 2010. **87**(2): p. 219-28.
377. Karamboulas, C., et al., *HDAC activity regulates entry of mesoderm cells into the cardiac muscle lineage.* J Cell Sci, 2006. **119**(Pt 20): p. 4305-14.
378. Yi, W., et al., *Phosphofructokinase 1 glycosylation regulates cell growth and metabolism.* Science, 2012. **337**(6097): p. 975-80.
379. Town, L., et al., *The metalloendopeptidase gene Pitrm1 is regulated by hedgehog signaling in the developing mouse limb and is expressed in muscle progenitors.* Dev Dyn, 2009. **238**(12): p. 3175-84.
380. Mittaz, L., et al., *Localization of a novel human RNA-editing deaminase (hRED2 or ADARB2) to chromosome 10p15.* Hum Genet, 1997. **100**(3-4): p. 398-400.
381. Sasman, A., et al., *Generation of conditional alleles for Foxc1 and Foxc2 in mice.* Genesis, 2012. **50**(10): p. 766-74.
382. Winnier, G.E., et al., *Roles for the winged helix transcription factors MF1 and MFH1 in cardiovascular development revealed by nonallelic noncomplementation of null alleles.* Dev Biol, 1999. **213**(2): p. 418-31.

383. Fuse, N., et al., *Novel mutations in the FOXC1 gene in Japanese patients with Axenfeld-Rieger syndrome.* Mol Vis, 2007. **13**: p. 1005-9.
384. Schouten, J.P., et al., *Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.* Nucleic Acids Res, 2002. **30**(12): p. e57.
385. Hofmann, J.J., et al., *Endothelial deletion of murine Jag1 leads to valve calcification and congenital heart defects associated with Alagille syndrome.* Development, 2012. **139**(23): p. 4449-60.
386. Krantz, I.D., et al., *Jagged1 mutations in patients ascertained with isolated congenital heart defects.* Am J Med Genet, 1999. **84**(1): p. 56-60.
387. Damert, A., et al., *Insufficient VEGFA activity in yolk sac endoderm compromises haematopoietic and endothelial differentiation.* Development, 2002. **129**(8): p. 1881-92.
388. Ferrara, N., et al., *Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene.* Nature, 1996. **380**(6573): p. 439-42.
389. Lambrechts, D., et al., *Low expression VEGF haplotype increases the risk for tetralogy of Fallot: a family based association study.* J Med Genet, 2005. **42**(6): p. 519-22.
390. Lui, T.T., et al., *The ubiquitin-specific protease USP34 regulates axin stability and Wnt/beta-catenin signaling.* Mol Cell Biol, 2011. **31**(10): p. 2053-65.
391. Chia, I.V., et al., *Both the RGS domain and the six C-terminal amino acids of mouse Axin are required for normal embryogenesis.* Genetics, 2009. **181**(4): p. 1359-68.
392. Hurlstone, A.F., et al., *The Wnt/beta-catenin pathway regulates cardiac valve formation.* Nature, 2003. **425**(6958): p. 633-7.
393. Lissitzky, J.C., et al., *Endoproteolytic processing of integrin pro-alpha subunits involves the redundant function of furin and proprotein convertase (PC) 5A, but not paired basic amino acid converting enzyme (PACE) 4, PC5B or PC7.* Biochem J, 2000. **346 Pt 1**: p. 133-8.
394. Szumska, D., et al., *VACTERL/caudal regression/Currarino syndrome-like malformations in mice with mutation in the proprotein convertase Pcsk5.* Genes Dev, 2008. **22**(11): p. 1465-77.
395. Pytela, R. and G. Wiche, *High molecular weight polypeptides (270,000-340,000) from cultured cells are related to hog brain microtubule-associated proteins but copurify with intermediate filaments.* Proc Natl Acad Sci U S A, 1980. **77**(8): p. 4808-12.
396. Natsuga, K., et al., *Plectin expression patterns determine two distinct subtypes of epidermolysis bullosa simplex.* Hum Mutat, 2010. **31**(3): p. 308-16.
397. Gundesli, H., et al., *Mutation in exon 1f of PLEC, leading to disruption of plectin isoform 1f, causes autosomal-recessive limb-girdle muscular dystrophy.* Am J Hum Genet, 2010. **87**(6): p. 834-41.
398. Konieczny, P., et al., *Myofiber integrity depends on desmin network targeting to Z-disks and costameres via distinct plectin isoforms.* J Cell Biol, 2008. **181**(4): p. 667-81.
399. Fukuda, M., et al., *Cloning of cDNAs encoding human lysosomal membrane glycoproteins, h-lamp-1 and h-lamp-2. Comparison of their deduced amino acid sequences.* J Biol Chem, 1988. **263**(35): p. 18920-8.

400. Nishino, I., et al., *Primary LAMP-2 deficiency causes X-linked vacuolar cardiomyopathy and myopathy (Danon disease).* Nature, 2000. **406**(6798): p. 906-10.
401. Arad, M., et al., *Glycogen storage diseases presenting as hypertrophic cardiomyopathy.* N Engl J Med, 2005. **352**(4): p. 362-72.
402. Charron, P., et al., *Danon's disease as a cause of hypertrophic cardiomyopathy: a systematic survey.* Heart, 2004. **90**(8): p. 842-6.
403. Tanaka, Y., et al., *Accumulation of autophagic vacuoles and cardiomyopathy in LAMP-2-deficient mice.* Nature, 2000. **406**(6798): p. 902-6.
404. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* Am J Hum Genet, 1993. **52**(3): p. 506-16.
405. Lewis, C.M., *Genetic association studies: design, analysis and interpretation.* Brief Bioinform, 2002. **3**(2): p. 146-53.
406. Ewens, W.J. and R.S. Spielman, *What is the significance of a significant TDT?* Hum Hered, 2005. **60**(4): p. 206-10.
407. LeClerc, S., et al., *Molecular cloning and characterization of a factor that binds the human glucocorticoid receptor gene and represses its expression.* J Biol Chem, 1991. **266**(26): p. 17333-40.
408. Brouns, M.R., et al., *The adhesion signaling molecule p190 RhoGAP is required for morphogenetic processes in neural development.* Development, 2000. **127**(22): p. 4891-903.
409. Arthur, W.T. and K. Burridge, *RhoA inactivation by p190RhoGAP regulates cell spreading and migration by promoting membrane protrusion and polarity.* Mol Biol Cell, 2001. **12**(9): p. 2711-20.
410. Kshitiz, et al., *Matrix rigidity controls endothelial differentiation and morphogenesis of cardiac precursors.* Sci Signal, 2012. **5**(227): p. ra41.
411. Goldenberg, I. and A.J. Moss, *Long QT syndrome.* J Am Coll Cardiol, 2008. **51**(24): p. 2291-300.
412. Westenskow, P., et al., *Compound mutations: a common cause of severe long-QT syndrome.* Circulation, 2004. **109**(15): p. 1834-41.
413. Tester, D.J., et al., *Compendium of cardiac channel mutations in 541 consecutive unrelated patients referred for long QT syndrome genetic testing.* Heart Rhythm, 2005. **2**(5): p. 507-17.
414. Millat, G., et al., *Spectrum of pathogenic mutations and associated polymorphisms in a cohort of 44 unrelated patients with long QT syndrome.* Clin Genet, 2006. **70**(3): p. 214-27.
415. Morimoto, S., *Sarcomeric proteins and inherited cardiomyopathies.* Cardiovasc Res, 2008. **77**(4): p. 659-66.
416. Martinsson, T., et al., *Autosomal dominant myopathy: missense mutation (Glu-706 --> Lys) in the myosin heavy chain IIa gene.* Proc Natl Acad Sci U S A, 2000. **97**(26): p. 14614-9.
417. Rutland, C.S., et al., *Knockdown of embryonic myosin heavy chain reveals an essential role in the morphology and function of the developing heart.* Development, 2011. **138**(18): p. 3955-66.
418. Kontrogianni-Konstantopoulos, A., et al., *Obscurin regulates the organization of myosin into A bands.* Am J Physiol Cell Physiol, 2004. **287**(1): p. C209-17.

419. Konstantopoulos, M.A.A.a.A.K.-. *Cardiomyopathies*, J.M.a.G. Ambrosio, Editor. 2013, InTech.
420. Svensson, E.C., et al., *Molecular cloning of FOG-2: a modulator of transcription factor GATA-4 in cardiomyocytes.* Proc Natl Acad Sci U S A, 1999. **96**(3): p. 956-61.
421. Tevosian, S.G., et al., *FOG-2, a cofactor for GATA transcription factors, is essential for heart morphogenesis and development of coronary vessels from epicardium.* Cell, 2000. **101**(7): p. 729-39.
422. Svensson, E.C., et al., *A syndrome of tricuspid atresia in mice with a targeted mutation of the gene encoding Fog-2.* Nat Genet, 2000. **25**(3): p. 353-6.
423. Hildebrand, J.D. and P. Soriano, *Overlapping and unique roles for C-terminal binding protein 1 (CtBP1) and CtBP2 during mouse development.* Mol Cell Biol, 2002. **22**(15): p. 5296-307.
424. Chen, J.D. and R.M. Evans, *A transcriptional co-repressor that interacts with nuclear hormone receptors.* Nature, 1995. **377**(6548): p. 454-7.
425. Jepsen, K., et al., *SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron.* Nature, 2007. **450**(7168): p. 415-9.
426. Vidal, O., et al., *Estrogen receptor specificity in the regulation of skeletal growth and maturation in male mice.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5474-9.
427. Arevalo, M.A., et al., *Estradiol meets notch signaling in developing neurons.* Front Endocrinol (Lausanne), 2011. **2**: p. 21.
428. Laherty, C.D., et al., *SAP30, a component of the mSin3 corepressor complex involved in N-CoR-mediated repression by specific transcription factors.* Mol Cell, 1998. **2**(1): p. 33-42.
429. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
430. UniProt, C., *Update on activities at the Universal Protein Resource (UniProt) in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D43-7.
431. Yuan, S., S. Zaidi, and M. Brueckner, *Congenital heart disease: emerging themes linking genetics and development.* Curr Opin Genet Dev, 2013. **23**(3): p. 352-9.
432. Gale, R.E., et al., *Acquired skewing of X-chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age.* Br J Haematol, 1997. **98**(3): p. 512-9.
433. McKellar, S.H., et al., *Novel NOTCH1 mutations in patients with bicuspid aortic valve disease and thoracic aortic aneurysms.* J Thorac Cardiovasc Surg, 2007. **134**(2): p. 290-6.
434. *Echocardiography in Pediatric and Adult Congenital Heart Disease*. 2012.
435. Craig E Fleishman, M.A.T., MD, *Clinical manifestations, pathophysiology, and diagnosis of atrioventricular (AV) canal defects*. UpToDate, ed. D.S. Basow. 2013: Waltham, MA.
436. Allen HD, S.R., Driscoll DJ, Feltes *Moss and Adams' Heart Disease in Infants, Children, and Adolescents Including the Fetus and Young Adult*. 2007, Lippincott Williams & Wilkins.

437. Rastelli, G., J.W. Kirklin, and J.L. Titus, *Anatomic observations on complete form of persistent common atrioventricular canal with special reference to atrioventricular valves.* Mayo Clin Proc, 1966. **41**(5): p. 296-308.
438. Reller, M.D., et al., *Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005.* J Pediatr, 2008. **153**(6): p. 807-13.
439. Hoffman, J.I., *Incidence of congenital heart disease: I. Postnatal incidence.* Pediatr Cardiol, 1995. **16**(3): p. 103-13.
440. Allan, L.D., et al., *Prospective diagnosis of 1,006 consecutive cases of congenital heart disease in the fetus.* J Am Coll Cardiol, 1994. **23**(6): p. 1452-8.
441. Peoples, W.M., J.H. Moller, and J.E. Edwards, *Polysplenia: a review of 146 cases.* Pediatr Cardiol, 1983. **4**(2): p. 129-37.
442. Services, T.D.o.S.H. *Texas Birth Defects Epidemiology and Surveillance*. 2011 20 July 2013]; Available from: http://www.dshs.state.tx.us/birthdefects/.
443. Agopian, A.J., et al., *Descriptive epidemiology of non-syndromic complete atrioventricular canal defects.* Paediatr Perinat Epidemiol, 2012. **26**(6): p. 515-24.
444. Rosenthal, G.L., et al., *Birth weight and cardiovascular malformations: a population-based study. The Baltimore-Washington Infant Study.* Am J Epidemiol, 1991. **133**(12): p. 1273-81.
445. Craig, B., *Atrioventricular septal defect: from fetus to adult.* Heart, 2006. **92**(12): p. 1879-85.
446. Calabro, R. and G. Limongelli, *Complete atrioventricular canal.* Orphanet J Rare Dis, 2006. **1**: p. 8.
447. Berger, T.J., et al., *Survival and probability of cure without and with operation in complete atrioventricular canal.* Ann Thorac Surg, 1979. **27**(2): p. 104-11.
448. Aubert, S., et al., *Atypical forms of isolated partial atrioventricular septal defect increase the risk of initial valve replacement and reoperation.* Eur J Cardiothorac Surg, 2005. **28**(2): p. 223-8.
449. Studer, M., et al., *Determinants of early and late results of repair of atrioventricular septal (canal) defects.* J Thorac Cardiovasc Surg, 1982. **84**(4): p. 523-42.
450. Abuhamad, A. and R. Chaoui, *A practical guide to fetal echocardiography : normal and abnormal hearts*. 2nd ed. 2010, Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins. vii, 379 p.
451. Eisenberg, L.M. and R.R. Markwald, *Molecular regulation of atrioventricular valvuloseptal morphogenesis.* Circ Res, 1995. **77**(1): p. 1-6.
452. Webb, S., N.A. Brown, and R.H. Anderson, *Formation of the atrioventricular septal structures in the normal mouse.* Circ Res, 1998. **82**(6): p. 645-56.
453. Snarr, B.S., C.B. Kern, and A. Wessels, *Origin and fate of cardiac mesenchyme.* Dev Dyn, 2008. **237**(10): p. 2804-19.
454. Snarr, B.S., et al., *Isl1 expression at the venous pole identifies a novel role for the second heart field in cardiac development.* Circ Res, 2007. **101**(10): p. 971-4.
455. Anderson, R.H., et al., *Development of the heart: (2) Septation of the atriums and ventricles.* Heart, 2003. **89**(8): p. 949-58.

456. Moorman, A., et al., *Development of the heart: (1) formation of the cardiac chambers and arterial trunks.* Heart, 2003. **89**(7): p. 806-14.
457. Patel, S.S., *Non-Syndromic atrioventricular septal defects: arefined definition, associated risk factors, and prognostic factors for left atrioventricular valve replacement following primary repair*, 2010, University of Iowa.
458. Carmi, R., J.A. Boughman, and C. Ferencz, *Endocardial cushion defect: further studies of "isolated" versus "syndromic" occurrence.* Am J Med Genet, 1992. **43**(3): p. 569-75.
459. Ferencz, C., et al., *Congenital cardiovascular malformations: questions on inheritance. Baltimore-Washington Infant Study Group.* J Am Coll Cardiol, 1989. **14**(3): p. 756-63.
460. Nemer, A.C.F.a.G.M., *Genetic Causes of Syndromic and Non-Syndromic Congenital Heart Disease*, in *Mutations in Human Genetic Disease*, P.D. Cooper, Editor. 2012, InTech.
461. Barlow, G.M., et al., *Down syndrome congenital heart disease: a narrowed region and a candidate gene.* Genet Med, 2001. **3**(2): p. 91-101.
462. Casas, C., et al., *Dscr1, a novel endogenous inhibitor of calcineurin signaling, is expressed in the primitive ventricle of the heart and during neurogenesis.* Mech Dev, 2001. **101**(1-2): p. 289-92.
463. Ackerman, C., et al., *An excess of deleterious variants in VEGF-A pathway genes in Down-syndrome-associated atrioventricular septal defects.* Am J Hum Genet, 2012. **91**(4): p. 646-59.
464. Green, E.K., et al., *Detailed mapping of a congenital heart disease gene in chromosome 3p25.* J Med Genet, 2000. **37**(8): p. 581-7.
465. Digilio, M.C., et al., *Atrioventricular canal and 8p- syndrome.* Am J Med Genet, 1993. **47**(3): p. 437-8.
466. Marino, B., et al., *Nonrandom association of atrioventricular canal and del (8p) syndrome.* Am J Med Genet, 1992. **42**(4): p. 424-7.
467. Mohapatra, B., et al., *Identification and functional characterization of NODAL rare variants in heterotaxy and isolated cardiovascular malformations.* Hum Mol Genet, 2009. **18**(5): p. 861-71.
468. Maslen, C.L., *Molecular genetics of atrioventricular septal defects.* Curr Opin Cardiol, 2004. **19**(3): p. 205-10.
469. O'Nuallain, S., J.G. Hall, and S.J. Stamm, *Autosomal dominant inheritance of endocardial cushion defect.* Birth Defects Orig Artic Ser, 1977. **13**(3A): p. 143-7.
470. Emanuel, R., et al., *Evidence of congenital heart disease in the offspring of parents with atrioventricular defects.* Br Heart J, 1983. **49**(2): p. 144-7.
471. Wilson, L., et al., *A large, dominant pedigree of atrioventricular septal defect (AVSD): exclusion from the Down syndrome critical region on chromosome 21.* Am J Hum Genet, 1993. **53**(6): p. 1262-8.
472. Kumar, A., C.A. Williams, and B.E. Victorica, *Familial atrioventricular septal defect: possible genetic mechanisms.* Br Heart J, 1994. **71**(1): p. 79-81.
473. Amati, F., et al., *Two pedigrees of autosomal dominant atrioventricular canal defect (AVCD): exclusion from the critical region on 8p.* Am J Med Genet, 1995. **57**(3): p. 483-8.

474. Cousineau, A.J., et al., *Linkage analysis of autosomal dominant atrioventricular canal defects: exclusion of chromosome 21.* Hum Genet, 1994. **93**(2): p. 103-8.
475. Robinson, S.W., et al., *Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects.* Am J Hum Genet, 2003. **72**(4): p. 1047-52.
476. Sheffield, V.C., et al., *Identification of a complex congenital heart defect susceptibility locus by using DNA pooling and shared segment analysis.* Hum Mol Genet, 1997. **6**(1): p. 117-21.
477. Phipps, M.E., et al., *Molecular genetic analysis of the 3p- syndrome.* Hum Mol Genet, 1994. **3**(6): p. 903-8.
478. Drumheller, T., et al., *Precise localisation of 3p25 breakpoints in four patients with the 3p-syndrome.* J Med Genet, 1996. **33**(10): p. 842-7.
479. Rupp, P.A., et al., *Identification, genomic organization and mRNA expression of CRELD1, the founding member of a unique family of matricellular proteins.* Gene, 2002. **293**(1-2): p. 47-57.
480. Guo, Y., et al., *Novel CRELD1 gene mutations in patients with atrioventricular septal defect.* World J Pediatr, 2010. **6**(4): p. 348-52.
481. Sarkozy, A., et al., *CRELD1 and GATA4 gene analysis in patients with nonsyndromic atrioventricular canal defects.* Am J Med Genet A, 2005. **139**(3): p. 236-8.
482. Zatyka, M., et al., *Analysis of CRELD1 as a candidate 3p25 atrioventicular septal defect locus (AVSD2).* Clin Genet, 2005. **67**(6): p. 526-8.
483. Smith, K.A., et al., *Dominant-negative ALK2 allele associates with congenital heart defects.* Circulation, 2009. **119**(24): p. 3062-9.
484. Rajagopal, S.K., et al., *Spectrum of heart disease associated with murine and human GATA4 mutation.* J Mol Cell Cardiol, 2007. **43**(6): p. 677-85.
485. Zhang, W., et al., *GATA4 mutations in 486 Chinese patients with congenital heart disease.* Eur J Med Genet, 2008. **51**(6): p. 527-35.
486. Maitra, M., et al., *Identification of GATA6 sequence variants in patients with congenital heart defects.* Pediatr Res, 2010. **68**(4): p. 281-5.
487. Stefansson, H., et al., *Large recurrent microdeletions associated with schizophrenia.* Nature, 2008. **455**(7210): p. 232-6.
488. Jun, G., et al., *Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.* Am J Hum Genet, 2012. **91**(5): p. 839-48.
489. Pearson, T.A. and T.A. Manolio, *How to interpret a genome-wide association study.* JAMA, 2008. **299**(11): p. 1335-44.
490. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
491. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-7.
492. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
493. Zhang, K., et al., *Molecular cloning and characterization of three novel lysozyme-like genes, predominantly expressed in the male reproductive system of humans, belonging to the c-type lysozyme/alpha-lactalbumin family.* Biol Reprod, 2005. **73**(5): p. 1064-71.

494. Richardson, L., et al., *EMAGE mouse embryo spatial gene expression database: 2010 update.* Nucleic Acids Res, 2010. **38**(Database issue): p. D703-9.
495. Hirano, S., et al., *Identification of a neural alpha-catenin as a key regulator of cadherin function and multicellular organization.* Cell, 1992. **70**(2): p. 293-301.
496. Cook, S.A., et al., *Cerebellar deficient folia (cdf): a new mutation on mouse chromosome 6.* Mamm Genome, 1997. **8**(2): p. 108-12.
497. Futterer, A., et al., *Ablation of Dido3 compromises lineage commitment of stem cells in vitro and during early embryonic development.* Cell Death Differ, 2012. **19**(1): p. 132-43.
498. Gronda, M., et al., *Hematopoietic protein tyrosine phosphatase suppresses extracellular stimulus-regulated kinase activation.* Mol Cell Biol, 2001. **21**(20): p. 6851-8.
499. Hentschke, M. and U. Borgmeyer, *Identification of PNRC2 and TLE1 as activation function-1 cofactors of the orphan nuclear receptor ERRgamma.* Biochem Biophys Res Commun, 2003. **312**(4): p. 975-82.
500. Fossey, S.C., et al., *Identification and characterization of PRKCBP1, a candidate RACK-like protein.* Mamm Genome, 2000. **11**(10): p. 919-25.
501. Lin, F.J., et al., *Endocardial cushion morphogenesis and coronary vessel development require chicken ovalbumin upstream promoter-transcription factor II.* Arterioscler Thromb Vasc Biol, 2012. **32**(11): p. e135-46.
502. Kruse, S.W., et al., *Identification of COUP-TFII orphan nuclear receptor as a retinoic acid-activated receptor.* PLoS Biol, 2008. **6**(9): p. e227.
503. Tsai, S.Y. and M.J. Tsai, *Chick ovalbumin upstream promoter-transcription factors (COUP-TFs): coming of age.* Endocr Rev, 1997. **18**(2): p. 229-40.
504. Winston, J.B., et al., *Heterogeneity of genetic modifiers ensures normal cardiac development.* Circulation, 2010. **121**(11): p. 1313-21.
505. Hardenbol, P., et al., *Multiplexed genotyping with sequence-tagged molecular inversion probes.* Nat Biotechnol, 2003. **21**(6): p. 673-8.
506. Schippers, A., et al., *Mucosal addressin cell-adhesion molecule-1 controls plasma-cell migration and function in the small intestine of mice.* Gastroenterology, 2009. **137**(3): p. 924-33.
507. de la Pompa, J.L., et al., *Role of the NF-ATc transcription factor in morphogenesis of cardiac valves and septum.* Nature, 1998. **392**(6672): p. 182-6.
508. Dor, Y., et al., *A novel role for VEGF in endocardial cushion formation and its potential contribution to congenital heart defects.* Development, 2001. **128**(9): p. 1531-8.
509. Digilio, M.C., et al., *Cardiac malformations in patients with oral-facial-skeletal syndromes: clinical similarities with heterotaxia.* Am J Med Genet, 1999. **84**(4): p. 350-6.
510. Ruiz-Perez, V.L., et al., *Evc is a positive mediator of Ihh-regulated bone growth that localises at the base of chondrocyte cilia.* Development, 2007. **134**(16): p. 2903-12.
511. Sund, K.L., et al., *Analysis of Ellis van Creveld syndrome gene products: implications for cardiovascular development and disease.* Hum Mol Genet, 2009. **18**(10): p. 1813-24.

512. Lin, F.J., et al., *Coup d'Etat: an orphan takes control.* Endocr Rev, 2011. **32**(3): p. 404-21.
513. Pereira, F.A., et al., *The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development.* Genes Dev, 1999. **13**(8): p. 1037-49.
514. Correa, A., et al., *Diabetes mellitus and birth defects.* Am J Obstet Gynecol, 2008. **199**(3): p. 237 e1-9.
515. Botto, L.D., et al., *Vitamin A and cardiac outflow tract defects.* Epidemiology, 2001. **12**(5): p. 491-6.
516. Perilhou, A., et al., *The transcription factor COUP-TFII is negatively regulated by insulin and glucose via Foxo1- and ChREBP-controlled pathways.* Mol Cell Biol, 2008. **28**(21): p. 6568-79.
517. Vilhais-Neto, G.C., et al., *Rere controls retinoic acid signalling and somite bilateral symmetry.* Nature, 2010. **463**(7283): p. 953-7.
518. Nakamura, E., et al., *5.78 Mb terminal deletion of chromosome 15q in a girl, evaluation of NR2F2 as candidate gene for congenital heart defects.* Eur J Med Genet, 2011. **54**(3): p. 354-6.
519. Zollner, S. and J.K. Pritchard, *Overcoming the winner's curse: estimating penetrance parameters from case-control data.* Am J Hum Genet, 2007. **80**(4): p. 605-15.
520. Ulucan, H., et al., *Extending the spectrum of Ellis van Creveld syndrome: a large family with a mild mutation in the EVC gene.* BMC Med Genet, 2008. **9**: p. 92.
521. Langheinrich, U., et al., *Zebrafish as a model organism for the identification and characterization of drugs and genes affecting p53 signaling.* Curr Biol, 2002. **12**(23): p. 2023-8.
522. Robu, M.E., et al., *p53 activation by knockdown technologies.* PLoS Genet, 2007. **3**(5): p. e78.
523. Bill, B.R., et al., *A primer for morpholino use in zebrafish.* Zebrafish, 2009. **6**(1): p. 69-77.
524. Staudt, D. and D. Stainier, *Uncovering the molecular and cellular mechanisms of heart development using the zebrafish.* Annu Rev Genet, 2012. **46**: p. 397-418.
525. Bakkers, J., *Zebrafish as a model to study cardiac development and human cardiac disease.* Cardiovasc Res, 2011. **91**(2): p. 279-88.
526. Shaner, N.C., et al., *Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein.* Nat Biotechnol, 2004. **22**(12): p. 1567-72.
527. Pipaon, C., S.Y. Tsai, and M.J. Tsai, *COUP-TF upregulates NGFI-A gene expression through an Sp1 binding site.* Mol Cell Biol, 1999. **19**(4): p. 2734-45.
528. Achatz, G., et al., *Functional domains of the human orphan receptor ARP-1/COUP-TFII involved in active repression and transrepression.* Mol Cell Biol, 1997. **17**(9): p. 4914-32.

Beside the genetic components required to support newly identified CHD genes in trios and case/control study designs, **functional experiments** are essential to confirm the pathogenic effect of genes in animal models using knockout or knockdown experiments in mouse and zebrafish models. Where appropriate, the pathogenic effect of specific variants can also be investigated using cell-based assays such as luciferase activity experiments. Moreover, integrating exome and genome sequence data with gene expression data using RNA-Seq from fetal heart tissues at different developmental stages are likely to be a helpful tool to prioritize candidate genes. Integrating high-throughput genetics, functional genomics and cellular and animal modeling will require concerted effort and collaboration.

Appendix A

The following work was perfomed by Sebastian Gerety as part of chapter 3.

Methods: Functional experiments

Morpholino oligonucleotides (MO) were purchased from Gene Tools (Oregon, USA). One- to four-cell embryos were microinjected with 1.8 nl of morpholino diluted in water. The sequences of morpholinos used were zmym2 MO1: CTGAGTGTGGATGAATTACCAGATC, zmym2 MO2: ATTAAAATGACGTACTTCTTGCACA and tp53 GCGCCATTGCTTTGCAAGAATTG [521]. To eliminate off-target effects of morpholinos [522] we co-injected zmym2 MOs with tp53 MO.

The efficacy of the splice-blocking zmym2 MO1 was tested by RT-PCR. Embryos were injected with zmym2 MO1 or control MO, and grown until 24 hpf. RNA was extracted, and subject to RT-PCR with exonic primers spanning the targeted splice site, to detect correctly spliced mRNA. Additionally, to detect increased unspliced RNA, the above exonic primer was paired with a downstream intronic primer.
PRIMERS:
ZMYM2 MO1 Forward: CAAAAGTGGCGCTCTACCGTCTC
ZMYM2 MO1 Reverse exonic: GACGCCGATTGGGAGATCCATG

Results: Zebrafish morpholino knockout experiments

To assess whether *ZMYM2* has a role in heart development, my colleague, Sebastian Gretey, chose to perform loss of function experiments in the Zebrafish *Danio Rerio*. Their rapid, external development and a near-transparent body combined with rapid antisense oligo-mediated loss of function permits us to analyse gene function without the need for complex knockout technology.

Using the Ensembl browser, he first identified the zebrafish orthologue of *ZMYM2*, also called zmym2 (ENSDARG00000027353). The predicted zebrafish protein has a 50% amino acid identity with human *ZMYM2*, and shared synteny between the two species. Using the ENSEMBL predicted intron/exon structure of the zebrafish gene, Sebastian designed two antisense morpholinos, targeting the splice site at the end of the first and second coding exons. Injection of either of these morpholinos is predicted to cause intron retention, leading to premature truncation of the zmym2 transcript [523].

To determine if the morpholinos are effective at blocking splicing, he performed RT-PCR on injected embryos, which confirmed that zmym2 morpholino#1 injected embryos have an increase in unspliced mRNA and a decrease in correctly spliced mRNA across the target region (see Methods). These data confirm that zmym2 morpholino injection should decrease Zmym2 protein expression in the zebrafish embryo.

During heart morphogenesis in the zebrafish, a centrally aligned linear heart tube undergoes a lateral movement termed 'jogging', positioning it on the left side of the body by 24 hours post fertilisation (hpf) [524, 525]. Subsequent looping events in the second 24 hours of development results in an S-shaped heart structure resembling other vertebrate embryonic hearts, with ongoing blood flow. A number of genes implicated in ToF are linked to left-right asymmetry. To see whether the developing hearts in zmym2 morpholino injected embryos display any morphological defects, including aberrant jogging of the heart tube, or subsequent heart looping, both of which are strongly dependent on left-right asymmetry, he stained the heart tissue of zmym2 or control morpholino injected embryos by in situ hybridization with a CMLC2 RNA probe.

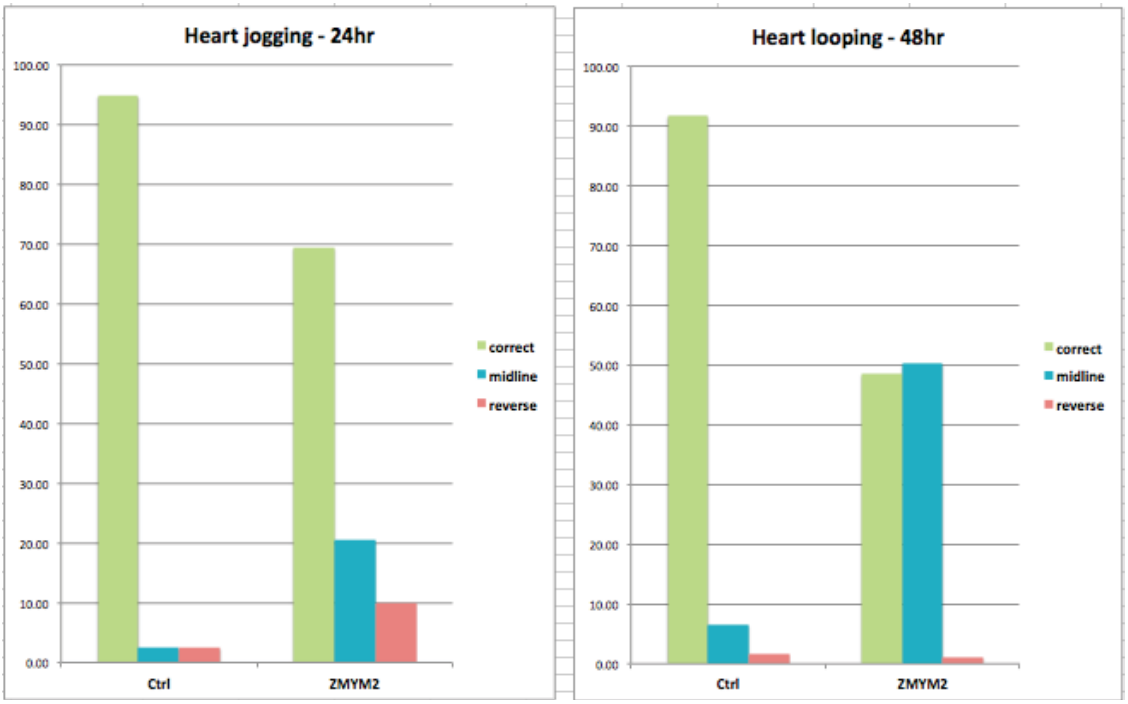


Figure A-1 Heart jogging and looping of the heart in the wild type and ZMYM2 morpholino injected embryos.

At 24 hpf, 94.9% of control injected embryos (n=91 embryos, 2 experiments) show a left jogging heart, while only 69.5% of zmym2 MO1 injected embryos show left jogging, with the remainder either central, or right jogging (n=70 embryos, 2 experiments). When analysed for heart looping at 48 hpf, 91.8% of control MO injected embryos (n=141, 3 experiments) showed correct looping, while only 48.5% of zmym2 MO1 injected embryos showed correct looping (Figure A-1). The remaining zmym2 MO1 injected embryos displayed a linear heart tube, in which looping had not occurred. The severity of the heart and other embryonic defects in zmym2 MO injected zebrfish results in dead or dying embryos by the fifth day post fertilisation.

Appendix B

The following work was perfomed by Sebastian Gerety as part of chapter 4.

Methods: *NR2F2* expression plasmids and luciferase constructs

My colleague, Sebastian Gerety, generated expression plasmids for *NR2F2* and its variants, the human wildtype *NR2F2* coding sequence was PCR amplified from a full length EST (Genbank acc.#BC042897), and cloned by Gibson assembly (New England Biolabs) into a CMV-driven pCS2-Cherry plasmid. To recreate the mutant forms of *NR2F2* (p.Lys70LysGln, p.Asp170Val, p.Asn205Ile, p.Glu251Asp, p.Ser341Tyr, and p.Ala412Ser), he amplified two PCR fragments overlapping each mutation, and cloned these as above. These expression constructs produce fusion proteins with fluorescent cherry domain [526] in order to monitor expression and localisation. To create the *NGFI-A* and *APOB* promoter driven Luciferase plasmids, he cloned synthetic DNA fragments for the rat *NGFI-A* upstream genomic region from -389 to +43 [527], and the human *APOB* upstream region from -139 to +121 [528], into a promoterless pGL3 Luciferase plasmid (Promega) by Gibson assembly (New England Biolabs).

Methods: Luciferase assays

HEK293T and HEPG2 cells were plated in 96-well plates, and transfected with 30ng of either *NGFI-A* or *APOB* luciferase plasmids, 0.75 ng of RL-TK renilla plasmid (Promega), and either 30ng of *NR2F2* expression plasmid (wildtype or variants) or 30ng of Cherry plasmid as a control. Two days after transfection, the cells were lysed and assayed for luciferase activity using the Dual-Luciferase Reporter Assay System, according to the manufacturer's instructions (Promega). Each transfection was done in replicates (minimum three times) and the experiments were repeated 3-4 times. Luciferase readings were first normalized to the transfection control (renilla plasmid). Relative Response Ratios (Promega) were calculated based on negative and positive controls (cherry and *NR2F2* plasmid transfections), and outliers across all experiments were

identified by a median absolute deviation ratio >3. A t-test was performed to identify significant differences between variants and between promoters.

Results: Luciferase assays

Despite the availability of computational methods predicting the effect of missense variants on protein function, interpreting the significance of these mutations in human disease is notoriously difficult. My colleague Sebastian Gerety tested the consequence of the identified *NR2F2* variants in a functional assay. Nr2f2 is a transcriptional regulator, with both activating and repressive effects on target gene expression [512]. A number of *NR2F2* responsive genomic elements have been identified, which when placed upstream of a reporter gene can quantitate transcriptional regulator function of Nr2f2 variants [502, 527, 528]. Using the most widely employed element, the promoter region of the *NGFI-A* gene [527], to drive a luciferase reporter in HEK293 cells, he compared its level of activation by wildtype *NR2F2* with that of the patient-derived variants. Sebastian observed robust luciferase activation by wildtype Nr2f2, and equivalent levels of activity from variants p.Asp170Val and p.Ala412Ser. However, two variants (p.Glu251Asp and p.Ser341Tyr) show a significantly lower activity in this assay (20-24% reduction, $p<0.01$), while variants p.Lys70LysGln and p.Asn205Ile have an increased activity (13-15% increase, $p<0.03$) (Figure B-1).

As the function of nuclear receptors involves a complex interaction with other transcriptional coregulators, he hypothesized that the consequence of Nr2f2 mutations might be promoter context dependent. Sebastian therefore performed the luciferase assay on an alternative promoter fragment from the *APOB* gene, that has previously been shown to be bound by Nr2f2 and used for structure-function studies [528]. In agreement with our prediction, the activities of the variants on the APOB promoter in HEK293 cells were significantly different from those using the *NGFI-A* promoter (Figure B-1). Variants p.Asp170Val, p.Asn205Ile, p.Glu251Asp and p.Ser341Tyr all show strong reductions in transcriptional activity compared to wildtype Nr2f2 (26-52%