Whole-Exome Capture and Next-Generation Sequencing to discover rare variants predisposing to congenital heart disease

Matthieu J. Miossec

For the degree of

Doctor of Philosophy

Newcastle University

Faculty of Medical Sciences

Institute of Genetic Medicine

December 2015

Newcastle University

Institute of Genetic Medicine

Abstract

Congenital heart disease (CHD) is the most common congenital malformation, affecting 8 out of 1000 lives births, yet its aetiology remains largely unresolved. The rapidly growing number of point mutations implicated in isolated CHD suggests that single mutations may contribute significantly to CHD risk. This thesis presents an investigation of the genetic underpinnings of various types of CHD following different study designs.

First, I designed a new approach to variant calling which I implemented as the variant caller BAMily. My aim was to develop a method of uncovering putative variants in next-generation sequencing data, shared by a subset of individuals and absent in another subset. I tested the variant caller's performance against other known variant callers and demonstrated that it provides comparable; and often better, results. This novel variant caller was applied to a study of 8 families in which a disease trait was segregating; along with the variant caller SAMtools, leading to the discovery of likely disease-causing variants in 5 families.

Second, I studied *de novo* mutation in 32 sporadic cases of transposition of the great arteries (TGA) in an attempt to identify genes that, when mutated, lead to TGA. The 32 patients with TGA were sequenced with their parents; as well as one unaffected sibling. To achieve this aim, three variant callers were used: SAMtools, GATK Unified Genotyper and BAMily, the latter acting as a filter. Potential *de novo* variants were found in *GREB1, RBP5, SNX13*. Results suggested a complex genetic etiology underlying TGA.

Finally, I studied a large series of cases of tetralogy of Fallot (ToF). The study involved 824 patients which ToF and a comparator set of 490 patients with neurodevelopmental disorders lifted from the UK10K project. The aim of the study was to identify genes that, when mutated, play a role in the manifestation of ToF might cluster. For this, I first categorised variants according to their potential to disrupt protein function. I then compared genes in which potentially disease-causing rare variants occurred to lists of genes previously implicated in CHD in the literature. Following this, I identified the clustering of potentially deleterious rare variants across the coding region of genes and exons in ToF patients, hypothesising that variants influencing ToF would cluster in ToF patients. This study led to the discovery of candidate variants in *FLT4* and *NOTCH1* for non-syndromic ToF. As with TGA, the results I have obtained suggested a complex etiology for ToF.

Acknowledgments

First and foremost, I would like to thank my supervisors Prof. Bernard Keavney, Dr. Mauro Santibáñez-Koref and Prof. Judith Goodship, for their excellent supervision. I consider myself fortunate to have had such dedicated supervisors. I would also like to thank Prof. Helen Arthur and Dr. Joana Elson for diligently reviewing my progress over the course of four years.

A special thanks goes to all those with whom I have had the pleasure of sharing an office, Prof. Heather Cordell, Dr. Ian Wilson, Holly Ainsworth, Kristin Ayers, Rebecca Darlay, Marla Endriga, Jakris 'Nat' Eu-Ahsunthornwattna, Osagie Ginikachukwu Izuogu, Helen Griffin, Darren Houniet, Richard Howey, Mikyung Jang, Michael Keogh, Valentina Mamasoula, Ali Qatan, Rachel Queen, So-Youn Shin, Wei Wei, Yaobo Xu. Thank you Darren for taking the time to show me the ropes in my first year. I would also like to thank the IT team; Arron Scott and Bryan Hepworth, for their timely responses to my many queries.

Over the course of four years, I have worked with a lot of talented people: Dr. Danielle Brown, Dr. Elise Glen, Dr. Darroch Hall, Mr. Rafiqul 'Raf' Hussain, Dr. Ruairidh Martin, Mr. Mzwandile Mbele, Thahira Rhaman, Dr. Louise Sutcliffe, Dr. Gennadiy Tenin, Dr. Ana Topf and Mrs. Helen Weatherstone. I have enjoyed working with each of you. Elise, Ana and Thahira have helped me tremendously over the years. Raf has always been quick to respond to questions and is an absolute pleasure to talk to. I have also had the privilege to supervise a great Masters student, Wirginia Bada, for which I am grateful.

I would also like to acknowledge my dear friend Laurence Dawson for encouraging me to pursue a doctorate.

Thank you to my family, my parents Trisha and Michel, without whom none of this would have been possible.

Finally, thank you to my beautiful and talented partner, Sandra Morales, for her endless support and encouragement. You have been by my side, through the good days and the bad days, and for this I am eternally grateful. I love you.

Statement of contribution

Unless specified otherwise, the work presented in this thesis is my own.

My focus in this thesis was the development of a novel bioinformatic tool for the analysis of next-generation sequencing data, and the analysis and interpretation of familial, trio and case-control datasets in congenital heart disease. Other aspects of the work were performed as outlined below.

Prof. Judith Goodship, Prof. Bernard Keavney and their collaborators; Prof. Bongani Mayosi, Prof. Paul Brennan and Dr. Andrew Harper, provided the samples and clinical information for the 8 familial cases studied in Chapter 3 and for the 32 families studied in Chapter 4. The ToF samples analysed in Chapter 5 were collected from 9 different locations across Europe and Australia as part of collaborative study coordinated by Prof. Bernard Keavney, Prof. Judith Goodship and Prof. Mark Lathrop; based at McGill University and Génome Québec Innovation Centre.

Sequencing performed at the Institute of Genetic Medicine in Newcastle was done under the supervision of Mr. Rafiqul Hussein, with sequencing data delivered by Bryan Hepworth. Sequencing for Chapter 4 and Chapter 5 was also performed by collaborators at Glasgow Polyomics, the McGill University and Génome Québec Innovation Centre, and The University of Manchester. Microarray and sequencing data for 21 parent-offspring trios used to test my variant caller in Chapter 3 were provided by Dr. Matthew Hurles.

Variant validations were performed by Dr. Elise Glen, Dr. Thahira Rahman, Dr. Ana Topf, Dr. Gennadiy Tenin and Dr. Louise Sutcliffe.

The method used to measure variant calling sensitivity and specificity in Chapter 5 was developed by Dr. Darren Houniet.

Table of Contents

List of figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Appendix

List of tables

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Appendix

Abbreviations

A2M	**Alpha-2-macroglobulin**
AAA	**Abdominal aortic aneurysm**
AD	**Autism spectrum disorder**
ADAMTS	**ADAM metallopeptidase with thrombospondin**
ANGPTL	**Angiopoietin-like**
ANK	**Ankyrin**
Ao	**Aorta**
AoS	**Aortic stenosis**
AoV	**Aortic valve**
APC	**Adenomatosis polyposis coli**
ASD	**Atrial septal defect**
ATG	**Autophagy related**
AVSD	**Atrioventricular septal defect**
BAV	**Bicuspid aortic Valve**
BF	**Bayes factor**
BLK	**Tyrosine kinase, B-lymphocyte-specific**
BVES	**Blood vessel epicardial substance**
CAST	**Cohort allelic sum test**
CCDC	**Coiled-coil domain containing**
CDCV	**Common disease, common variants**
CDRV	**Common disease, rare variants**
CDS	**Coding sequence**
CREB	**CAMP responsive element binding protein**
CFC1	**Cripto, frl-1, cryptic family**
CHD	**Congenital heart disease**
CHD7	**Chromodomain Helicase DNA Binding Protein 7**
CLIP	**CAP-GLY Domain Containing Linker Protein Family**
CLTCL	**Clathrin, heavy chain-like 1**
CMC	**Combined and multivariate and collapsing**
CNV	**Copy number variant**
CoA	**Coarctation of the aorta**
COL	**Collagen**
CPT	**Carnitine palmitoyltransferase**
CS	**Combined score**
CTBP	**C-terminal binding protein**
CTPS	**CTP synthase**

DCDC	**Doublecortin domain containing**
DCG3	**Discs, large homolog (drosophila)**
DCRS	**Danish civil registration system**
DOCK	**Dedicator of cytokinesis**
DORV	**Double-outlet right ventricle**
DMXL	**Dmx-like protein**
DNAH	**Dynein, axonemal, heavy chain**
D-TGA	**Dextro-transposition of the great arteries**
ECMO	**Extracorporeal membrane oxygenation**
EGA	**European genome-phenome archive**
ENTPD	**Ectonucleoside triphosphate diphosphohydrolase**
EVS	**Exome variant server**
EXOG	**Endo/Exonuclease, Endonuclease G-Like**
FLT	**FMS-like tyrosine kinase**
FOX	**Forkhead box**
GATA	**GATA binding protein**
GCKR	**Glucokinase (hexokinase 4) regulator**
GDF	**Growth differentiation factor**
GDPGP	**GDP-D-glucose phosphorylase**
GHR	**Growth hormone receptor**
GJA	**Gap junction protein, alpha**
GO	**Gene ontology**
GP	**Glasgow Polyomics**
GPC	**Glypican**
GPR	**G protein-coupled receptor**
GRC	**Genome reference consortium**
GREB	**Growth regulation by estrogen in breast cancer**
GRIK	**Glutamate receptor, ionotropic, kainate**
GWAS	**Genome-wide association study**
HAND	**Heart and neural crest derivatives**
HECTD4	**HECT domain containing E3 ubiquitin protein ligase**
HES	**Hairy gene homolog (drosophila)**
HEY	**Hes-related family BHLH transcription factor**
HGP	**Human genome project**
HMCN	**Hemicentin**
HRC	**Heart Repair Consortium**
HSP	**Hereditary scleroscing poikiloderma**
IAA	**Interrupted aortic arch**

INDEL	**Insertion/deletion**
IGF	**Insulin-Like Growth Factor**
IGM	**Institute of Genomic Medicine**
IGV	**Ingenuity variant analysis**
IVC	**Inferior vena cava**
JAG	**Jagged**
KCNH	**Potassium channel, voltage gated eag related subfamily H**
KCNJ	**Potassium channel, inwardly rectifying subfamily j**
KIR	**Killer cell immunoglobin-like receptor**
KRTAP	**Keratin associated protein**
KWE	**Keratolytic winter erythmea**
LA	**Left atrium**
LARS	**Leucyl-TRNA synthetase**
LMOD	**Leiomodin**
L-TGA	**Levo-transposition of the great arteries**
LV	**Left ventricle**
MAF	**Minor allele frequency**
MAPCA	**Major aortopulmonary collateral artery**
MEF	**Myocyte enhancer factor**
MGA	**Malposition of the great arteries**
MGI	**Mouse Genome Informatics**
MPA	**Main pulmonary artery**
MUGQIC	**McGill University and Génome Québec Innovation Centre**
MV	**Mitral valve**
MYH	**Myosin, heavy chain**
MYO	**Myosin**
MYOM	**Myomesin**
NAA	**N(alpha)-acetyltransferase**
NAV	**Neuron navigator**
NHLBI ESP	**National Heart, Lung, Blood Institute Exome Sequencing Project**
NDD	**Neurodevelopmental disorders**
NEB	**Nebulin**
NELL	**Neural EGFL-like**
NGS	**Next-generation sequencing**
NKX2	**NK2 Homeobox**
NMNAT	**Nicotinamide nucleotide adenylyltransferase**

NODAL	**Nodal growth differentiation factor**
NOS	**Nitric oxide synthase**
NOTCH	**Notch gene homolog (drosophila)**
PA	**Pulmonary atresia**
PCDH	**Protocadherin**
PCR	**Polymerase chain reaction**
PDA	**Patent ductus arteriosus**
PDE	**Phosphodiesterase**
PDX	**Pancreatic and duodenal homeobox**
PHLPP	**PH domain and leucine rich repeat protein phosphatase**
POPDC	**Popeye domain containing**
PRRC	**Proline-rich coiled-coil**
PROK	**Prokineticin**
PS	**Pulmonary stenosis**
PTPN	**Protein tyrosine phosphatase, non-receptor**
PUS	**Pseudouridylate synthase**
PV	**Pulmonary valve**
RA	**Right atrium**
RAB	**Ras-relatedGTP-binding protein**
RAD52	**Rad52 gene homolog (Saccharomyces cerevisiae)**
RBP	**Retinol binding protein**
RV	**Right ventricle**
RVH	**Right ventricular hypertrophy**
RVOT	**Right Ventricular outflow tract**
RVS	**Robust variance score**
SCN	**Sodium channel, voltage gated**
SERBP	**SERPINE1 MRNA binding protein**
SETBP	**SET-binding protein**
SKAT	**Sequence kernel association test**
SLC	**Solute carrier**
SLIT	**Slit gene homolog (drosophila)**
SMAD	**Sma- and Mad-related protein**
SMRT	**Single molecule, real time**
SNCAIP	**Synuclein, alpha interacting protein**
SNP	**Single nucleotide polymorphism**
SNV	**Single nucleotide variant**
SNX	**Sorting nexin**
STAT	**Signal transducer and activator of transcription**
STX	**Syntaxin**
SVC	**Superior vena cava**

TANC	**Tetratricopeptide repeat, ankyrin repeat and coiled-coil containing**
TAVR	**Total anomalous venous return**
TBX	**T-Box**
TCEB	**Transcription elongation factor B**
TDGF	**Teratocarcinoma-derived growth factor**
TGA	**Transposition of the great arteries**
TGFB	**Transforming growth factor, Beta**
ToF	**Tetralogy of Fallot**
TPO	**Thyroid peroxidase**
TV	**Tricuspid valve**
VAD	**Ventricular assist device**
VEGF	**Vascular endothelial growth factor**
UCSC	**University of California Santa Cruz**
VCL	**Vinculin**
VSD	**Ventricular septal defect**
WES	**Whole-exome sequencing**
WGS	**Whole-genome sequencing**
WST	**Weighted sum test**
WTCCC	**Welcome Trust Case Control Consortium**
ZFHX3	**Zinc finger homeobox**
ZFPM2	**Zinc finger protein, FOG family member**
ZIC	**Zinc finger protein of cerebellum**
ZNF	**Zinc finger**

Chapter 1. Introduction

1.1. Summary

This thesis presents multiple whole-exome sequencing and capture (WES) studies performed on patients with congenital heart disease (CHD). It also describes a new approach to variant calling, a key step of sequence analysis. This introductory chapter provides an overview of CHD and its causes, paying particular attention to the contribution of genetic factors, including results of previous WES studies. This chapter also describes the steps involved in a sequencing study, with a particular emphasis on WES studies. The last section provides an overview of different sequencing study designs.

1.2. Congenital Heart Disease Overview

1.2.1. Definition and incidence

Congenital heart disease (CHD) in its broadest sense refers to heart malformations arising during cardiogenesis (Bruneau and Srivastava, 2014). Disruption of regulatory mechanisms acting on specific cell lineages in the maturing heart can lead to structural deviations from the heart's typical morphology, resulting in varying degrees of morbidity at birth and later in life (Bruneau, 2008).

A recurring figure given for the incidence of CHD is 8 per 1000 live births (Bernier *et al.*, 2010). As such, CHD accounts for close to a third of all major congenital defects (van der Linde *et al.*, 2011). However, this figure is based on the aggregation of the results from epidemiological studies using different criteria as to what constitutes CHD. One early study of CHD by Mitchell *et al.* (1971) set the requirement that a heart defect should be "actually or potentially of functional significance" to be considered a CHD. This definition excludes defects largely considered benign such as persistent left superior vena cava (Hoffman and Kaplan, 2002). One notable heart defect that is consequently largely excluded from most epidemiological studies is bicuspid aortic valve (BAV) as clinical symptoms typically do not manifest before 40 years of age (Hoffman and Kaplan, 2002). Factoring BAV in the incidence of CHD would put the total CHD incidence at 21 per 1000 live births (Hoffman and Kaplan, 2002). Other limiting factors in estimating CHD are regional variations in incidence and the effectiveness of diagnosis (Bernier *et al.*, 2010).

CHD can take on many distinct forms. The 8 most common forms of CHD among modest to severe cases are shown in Figure 1 with their reported birth prevalence by continent.

They are as follows: ventricular septal defect (VSD), atrial septal defect (ASD), tetralogy of Fallot (ToF), pulmonary stenosis (PS), aortic stenosis (AoS), patent ductus arteriosus (PDA), coarctation of the aorta (CoA) and transposition of the great arteries (TGA) (van der Linde *et al.*, 2011).

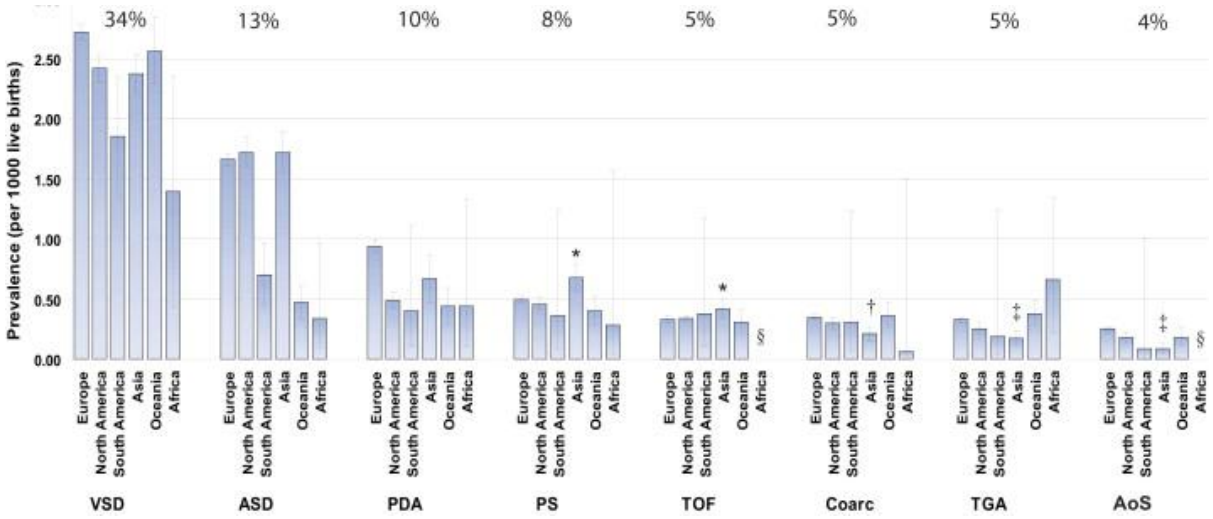

Figure 1. Birth prevalence of the 8 most common subtypes of modest to severe forms of CHD as measured on different continents. This bar chart is from van der Linde (2011) and is based on their meta-analysis of worldwide birth prevalence of CHD. Above each CHD subtype, the average prevalence is given. *Prevalence rate significantly higher than in Europe and North America. †Prevalence rate significantly lower than in Europe. ‡ Prevalence rate significantly lower than in Europe, North America and Oceania. § No data available.

Among CHD subtypes, VSD is the most common form of CHD. As an isolated cardiac defect, VSD accounts for nearly 40% of all diagnosed CHD (Penny and Vick, 2011). VSD also appears in complex forms of CHD, either as an integral part of a more complex condition; such as ToF, or as a secondary consequence; as in TGA. A VSD is a defect in the wall separating the heart's ventricles (Kung and Wong, 2014). The main outcome of VSD is the redirection of some of the heart's blood flow from one chamber directly into another (Penny and Vick, 2011). This bypassing of the circulatory system is referred to as 'shunting' and can occur in either or both directions, with varying levels of intensity, depending on the size and location of the defect. In some cases, the direction in which shunting occurs changes over time. Patients described as having Eisenmenger syndrome originally experience left-to-right shunting which, left untreated, turns into right-to-left shunting in young adult life owing to the development of pulmonary hypertension (Gamss and Haramati, 2014). In general, right-to-left shunting reduces blood flow to the lungs, thus limiting blood oxygenation (Penny and Vick, 2011). The mixing of deoxygenated blood with oxygenated blood manifests in patients as cyanosis, a bluish coloration of the skin (Bruneau, 2008). VSDs can also subsequently cause aortic valve prolapse or pulmonary valve obstruction (Penny and Vick, 2011). Most VSDs detected at birth are asymptomatic muscular defects that close within the first year of life (Penny and Vick, 2011). VSDs presenting a risk for a patient's health can be closed through surgery or the implantation of a catheter device (Kung and Wong, 2014).

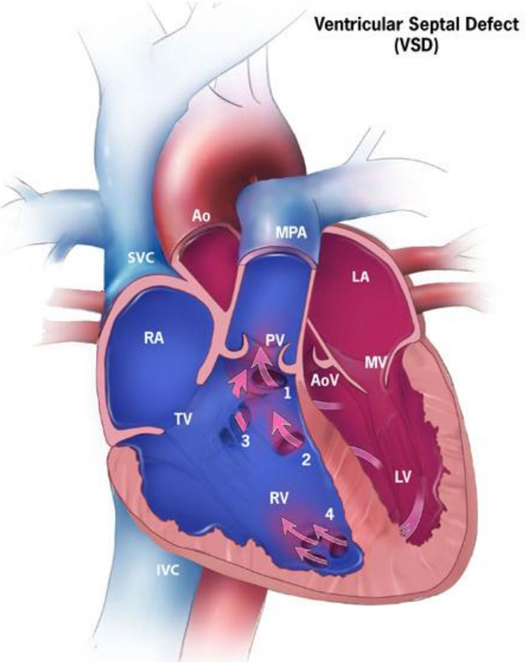

Figure 2. There are different types of ventricular septal defect (VSD) including conoventricular (1), perimembranous (2), inlet (3) and Muscular (4) VSD. The arrows represent the redirection of blood flow; also known as shunting. Abbreviations: RA: right atrium, RV: right ventricle, LA: left atrium, LV: left ventricle, SVC: superior vena cava, IVC: inferior vena cava, MPA: main pulmonary artery, Ao: aorta, TV: tricuspid valve, MV: mitral valve, PV: pulmonary valve, AoV: aortic valve. Image from the Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities available at: http://www.cdc.gov/ncbddd/heartdefects/ventricularseptaldefect.html

Defects in the septum may also arise at the level of the atria. ASDs are the second most common form of CHD totalling nearly 13% of all CHDs (Kendall *et al.*, 2014). ASD are further classified according to the atrial structure affected: the septum primum or the septum secundum (Kendall *et al.*, 2014). Septal defects in this latter structure, termed ostium secundum, are the most common, representing 80% of all ASDs (Kendall *et al.*, 2014). As with VSD, shunting occurs and its intensity is dependent on the location and size of the defect. Most ASDs requiring intervention are closed through the use of a catheter device with only extreme cases requiring surgery (Kendall *et al.*, 2014). An illustration of ASD is shown in Figure 3.

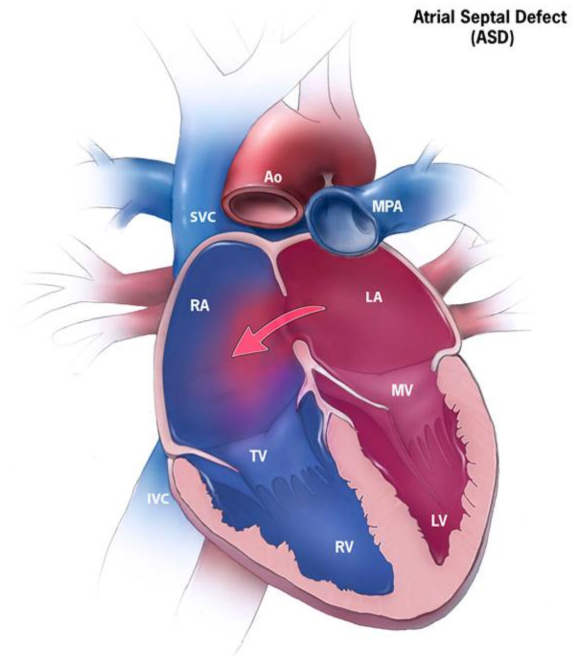

Figure 3. Atrial septal defect (ASD). The arrow crossing the septal defect represents the redirection of oxygenated blood flow from the left atrium to the right atrium. Abbreviations: RA: right atrium, RV: right ventricle, LA: left atrium, LV: left ventricle, SVC: superior vena cava, IVC: inferior vena cava, MPA: main pulmonary artery, Ao : aorta, TV: tricuspid valve, MV: mitral valve. Image from the Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities available at: http://www.cdc.gov/ncbddd/heartdefects/atrialseptaldefect.html

Another common type of cardiac abnormality involves the narrowing of structures of the heart, impeding blood flow. The three most common subtypes of CHD that fit this description are CoA, AoS and PS. According to Lee *et al*. (2014a), CoA is estimated to amount to 7% of all CHD cases, with a slight bias towards male cases. Individuals with CoA are likely to be diagnosed with other cardiac defects, particularly BAV (Lee *et al.*, 2014a). CoA describes the constriction of the stretch of aortic arch found between the left subclavian artery and the ductus arteriosus in the developing heart. Different levels of coarctation exist. Lee *et al.* (2014a) liken severe cases of CoA to cases of interrupted aortic arch as blood circulation through the aorta is blocked. On the other hand, some cases are asymptomatic. The latter are left mostly untreated (Lee *et al.*, 2014a). Both surgery and transcatheter balloon angioplasty have been extensively used to repair CoA with extremely low levels of morbidity and mortality in either case (Vergales *et al.*, 2013; Lee *et al.*, 2014a). Despite this, neither technique is fully restorative and patients are at risk of recoarctation, developing aneurysm or hypertension (Vergales *et al.*, 2013). As a result, lifelong follow-ups are necessary (Vergales *et al.*, 2013).

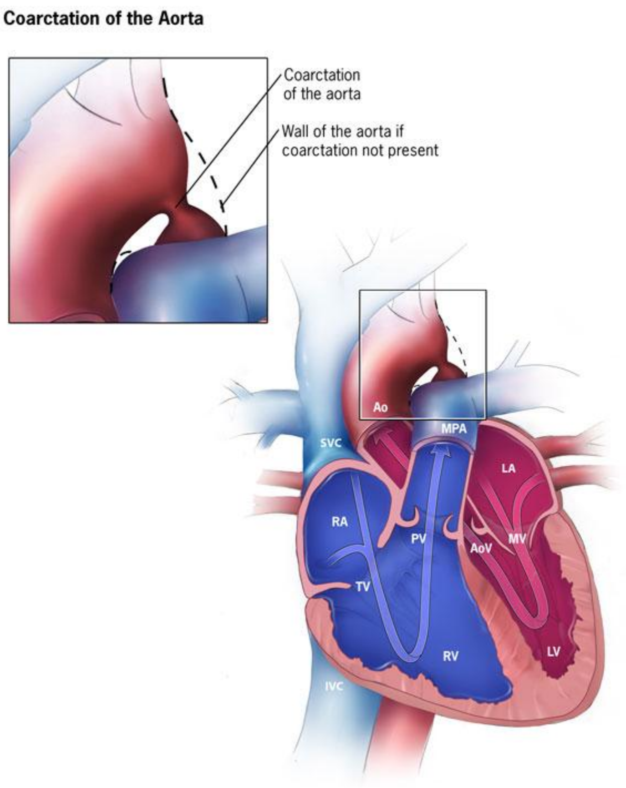

Figure 4. Coarctation of the aorta (CoA). The arrows show the direction of regular blood flow. Blood flow is impeded towards the descending aorta. Abbreviations: RA: right atrium, RV: right ventricle, LA: left atrium, LV: left ventricle, SVC: superior vena cava, IVC: inferior vena cava, MPA: main pulmonary artery, Ao: aorta, TV: tricuspid valve, MV: mitral valve, PV: pulmonary valve, AoV: aortic valve. Image from the Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities available at: http://www.cdc.gov/ncbddd/heartdefects/coarctationofaorta.html

AoS accounts for about 4% of all CHD cases (Hochstrasser *et al.*, 2014). AoS takes on the form of a narrowing of the aortic valve, called stenosis, which hinders blood flow from the left ventricle to the aorta (Alizadehasl and Sadeghpour, 2014). The stenosis is the result of the joining; termed commissural fusion, of normally distinct heart valve leaflets. The aortic valve is thus described in many cases as uniscupid; in predominantly severe cases, or bicuspid, although in some cases remains tricuspid (Alizadehasl and Sadeghpour, 2014). The additional resistance created by the stenosis puts a strain on the muscles of the heart's left chamber, leading to the development of an excessive, and eventually less efficient, left ventricle muscle mass. Balloon aortic valvuloplasty is performed on infants with the condition in order to widen the aortic valve with more invasive surgery taking place later in life (Ewert *et al.*, 2011; Alizadehasl and Sadeghpour, 2014). Long term treatment, particularly if stenosis reoccurs, includes aortic valve replacement either by inserting a mechanical aortic valve or grafting a donor's aortic or pulmonary valve (Alizadehasl and Sadeghpour, 2014).

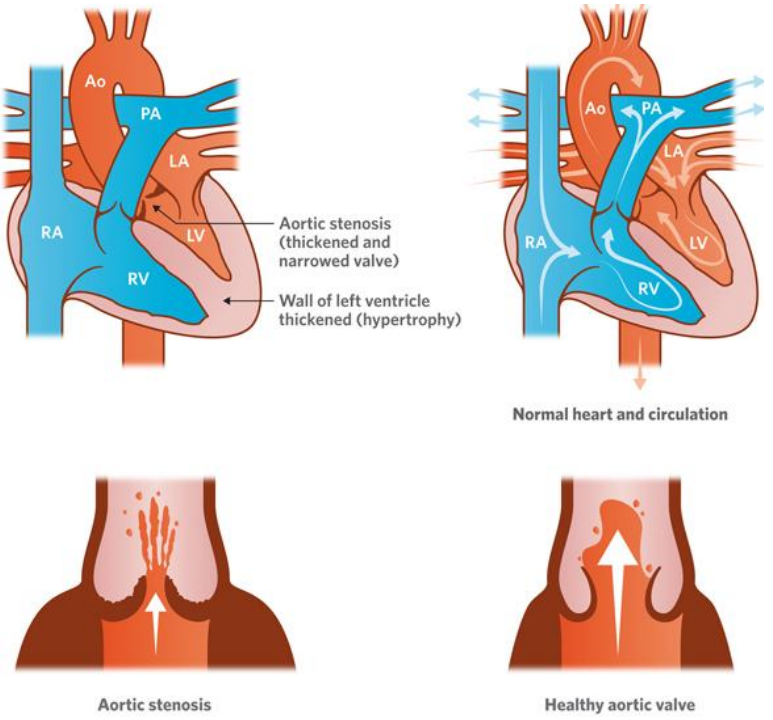
Ao
PA
LA
RA
LV
RV
Aortic stenosis (thickened and narrowed valve)
Wall of left ventricle thickened (hypertrophy)
Ao
PA
LA
RA
LV
RV
Normal heart and circulation
Aortic stenosis
Healthy aortic valve

Figure 5. Aortic stenosis (Aos). The arrows represent blood flow. In AoS, the blood flow to the aorta is limited by the partial fusion of valve leaflets. AoS can lead ventricular hypertrophy as shown above. Abbreviations: RA: right atrium, RV: right ventricle, LA: left atrium, LV: left ventricle, Ao: aorta, PA: pulmonary artery. Image is from The Royal Childent's Hospital Melbourne and is available at: http://www.rch.org.au/cardiology/heart_defects/Aortic_Stenosis_AS/

PS occurs in 7-12% of patients with CHD (Warnes *et al.*, 2008). PS is an obstructions similar to AoS, but located instead on the pulmonary valve (Sadeghpour and Alizadehasl, 2014). In this instance, the stenosis limits blood outflow towards the lungs. As with AoS, PS is the result of varying levels of commissural fusion (Sadeghpour and Alizadehasl, 2014). Depending on the severity of the PS, the right ventricle can display varying levels of hypertrophy and occasionally mild hypoplasia (Sadeghpour and Alizadehasl, 2014).

During foetal development, the aorta and the pulmonary artery are connected by the ductus arteriosus. This bridge usually disappears within two days of birth (Schneider and Moore, 2006). The ductus arteriosus is considered patent when the structure persists long after birth. For Schneider and Moore (2006), the structure can be considered patent if it persists beyond the first weeks after birth. However, Schneider and Moore (2006) acknowledge that other experts place the threshold for abnormality at 3 months. Excluding silent PDAs that are unintentionally discovered during echocardiography, PDA represents between 5 to 10% of all CHD cases (Schneider and Moore, 2006). PDAs lead to left-to-right shunting, some of the oxygenated blood returning to the pulmonary artery. The excess pulmonary blood flow results in an increase of pressure in the pulmonary circulation and may result in difficulties breathing (Schneider and Moore, 2006). If shunting is substantial and uncorrected, the rise in pulmonary pressure may achieve levels greater than the systemic circulation and become permanent, leading to Eisenmenger syndrome (Schneider and Moore, 2006). There is an active debate as to whether PDAs should be treated in the early years of life, when the rate of spontaneous ductus arteriosus closure is high, and if so, what treatments are preferable (Bose and Laughon, 2007; Clyman and Chorne, 2007).

ToF accounts for around 7-10% of CHD patients and is the most common cyanotic CHD (Bailliard and Anderson, 2009). The underlying abnormality in developmental anatomy resulting in ToF is typically considered to be an anterocephalad deviation of the outflow tract septum, which leads to an aggregation of four heart defects: a large VSD, a displacement of the aorta over the VSD, a PS causing the obstruction of the right ventricular outflow tract (RVOT), and hypertrophy of the right ventricle (Nelson *et al.*, 2014). This tetrad of abnormalities can be accompanied by other defects such as a right aortic arch or additional VSDs (Bailliard and Anderson, 2009). As such, ToF is one of the most complex forms of CHD. Surgery is performed in the first few months of life in order to close the VSD and mitigate the effect of the PS using the techniques usually associated with both these malformations in their isolated form (Mazur *et al.*, 2013). A more detailed description of ToF is presented in chapter 5.

TGA represents 5% of all CHD and is more prevalent in males (Unolt *et al.*, 2013). It is considered one of the most severe forms of CHD, carrying a 95% mortality rate within the first year of life if left untreated (Saremi, 2014). Patients with the most common form of TGA, dextro-transposition of the great arteries (D-TGA), display a pulmonary artery and an aorta that are each connected to the wrong ventricle (Warnes, 2006). Warnes (2006) describes this severe malformation as ventriculoarterial discordance. This discordance results in a lack of oxygenation of the blood travelling through the body as the circulation consists of two parallel circuits. Most cases are accompanied by additional cardiac malformations such as VSD, PS and other valvular abnormalies (Warnes, 2006). In the absence of a communication between the right and left circulations, the condition is lethal; therefore, initial intervention in TGA, in the early hours to days of life, typically consists in creating communication between the circulations at atrial level using balloon septostomy until a definite repair can be performed (Warnes, 2006). Where a complete

TGA is observed, an arterial switch operation is necessary (Warnes, 2006). A further description of TGA is given in chapter 4.

1.2.2. Evolution of treatment

While CHD remains the most prevalent cause of childhood morbidity in the Western world to this day, advances in medical and surgical care over the last few decades have made it possible for many of those afflicted to live beyond childhood (van der Linde *et al.*, 2011). Since the year 2000, more adults live with CHD than children (Guleserian, 2011). This has shifted some of the mortality from birth to adulthood.

The first serious attempt to correct a heart defect, was undertaken by John Strieder in 1937 (Kaemmerer *et al.*, 2004). However, the operation, which consisted in the ligation of a PDA, ended in failure, with the patient dying from complications 4 days later. The following year brought two successful PDA ligations, independently performed by Robert Gross and Emil Karl Frey within months of each other (Kaemmerer *et al.*, 2004). A few years later, in 1944, Clarence Crafoord performed the first successful CoA reparation (Guleserian, 2011). The same year, Blalock performed what was to be known as the Blalock-Taussig shunt operation (Guleserian, 2011). Designed to relieve the lack of blood oxygenation that characterises cyanotic heart disease, the operation redirects blood flow from the subclavian artery to the pulmonary artery. Many cardiac surgery firsts followed suit.

One important milestone in the treatment of heart defects is the development of the cardiopulmonary bypass unit, which allowed for effective open heart surgery (Guleserian, 2011). Previous surgeries of the same kind relied on hypothermia, presenting great risks for patients who did not receive blood to vital organs such as the brain (Guleserian, 2011). Cardiopulmonary bypass largely solved this issue. For Kendall *et al*. (2014), the first successful intracardial operation to correct ASD, performed by John Heysham Gibbon in 1953, marks the birth of modern cardiac surgery. Intracardial repair of multiple cardiac pathologies soon followed, with the successful operation of patients with VSD and ToF occurring less than a year after the first pioneering ASD surgery (Geva *et al.*, 2014).

The further development of heart surgery over a period of half a century has dramatically reduced CHD-related child mortality. Hoffman et al. (2004) estimate that in the US alone, 1.5 million patients will have been diagnosed with some form of CHD between 1940 and

2002. Conducting a large review of reports on various CHDs they concluded that, without surgery, we could expect 650,000 CHD patients to be alive in that time frame against more than 1.3 million, assuming equal access to available treatment. The actual number of CHD survivors is therefore somewhere in between these two estimates.

The increasing prevalence of CHD results chiefly from the ever-increasing number of patients surviving into adulthood (Hoffman *et al.*, 2004). This positive development comes with a caveat, that while heart surgery allows many CHD sufferers to survive beyond childhood, the surgery is rarely a fully restorative process. Operations on the heart leave behind what Perloff and Warnes (2001) describe as residua of the condition. In a patient with supravalvular aortic stenosis, systemic hypertension may persist even after surgical repair as some arterial abnormalities remain (Perloff and Warnes, 2001). Additionally, some surgical intervention, such as those which involve the grafting of cardiac structures, will inevitably inflict sequelae to the heart (Perloff and Warnes, 2001). For these reasons, adult survivors display a lower life expectancy when compared to that of the average population (Verheugt *et al.*, 2010). It is crucial for adults with CHD to receive proper medical advice on potential complications of corrected heart defects throughout life.

The shift in mortality rates caused by CHD from a primarily neonate population to adulthood also implies an additional disease burden on future generations where genetic variation plays a significant role in the disease (Fesslova *et al.*, 2011). In this context, genetic counselling becomes a crucial part of patient care. For prospective parents with a history of CHD risk, knowing the recurrence risk of a specific type of CHD in offspring allows them to make informed family planning decisions (Deanfield *et al.*, 2003). It also allows medical practitioners to better accompany at risk pregnancies (Deanfield *et al.*, 2003). Beyond recurrence risk, genetic testing, based on the rapidly evolving field of CHD genetics research, can provide a more targeted diagnosis for families (Pierpont *et al.*, 2007). Providing the most complete and up-to-date genetic counselling requires the identification of the whole range of genetic factors that contribute to CHD (Zaidi *et al.*, 2013).

1.2.3. Genetic contribution to CHD

In a seminal paper on the etiology of CHD, James Nora (1968) formulated the hypothesis that CHD was largely the result of a multitude of gene-environment interactions. Through this hypothesis, Nora suggests that cases of CHD are predominantly isolated events, with no single leading environmental or genetic cause (Gelb and Chung, 2014). However, successive population studies of CHD lead to a different conclusion, namely that genetic factors play a much larger role in CHD than would be expected if CHD were mainly the result of combined polygenic inheritance and environmental factors (Gelb and Chung, 2014). A case-control study of 2,102 infants with CHD in the Baltimore-Washington metropolitan area (Ferencz *et al.*, 1989) revealed a higher frequency of extracardiac abnormalities in CHD patients compared to healthy controls in the same area. 17.5% of infants with CHD had extracardiac abnormalities with an identified genetic cause, compared to 0.7% of healthy controls. From the CHD history of first degree relatives of 1893 cases, Ferencz *et al.* (1989) identified a significant excess of CHD in mothers and full siblings in cases compared with controls, further emphasizing the role inheritance can play in CHD. Burn *et al.* (1998) studied 727 adults with severe CHD, attempting a comprehensive nationwide ascertainment of the first cohort of CHD patients in the UK that had cardiac surgery in childhood and had at the time of the study started having families of their own. This study aimed to provide the recurrence risk in offspring for all CHD as well as distinct subtypes. Recurrence risk provides some insight into the fraction of a disease that is attributable to inherited risk. For adults with CHD, Burn *et al.* (1998) found a recurrence risk of CHD in offspring of 4.1%. This represents a fivefold higher prevalence than the 0.8% consensus estimate in an unselected population. However, this risk estimate conceals the strong variability that exists between CHD subtypes. Burn *et al.* (1998) report the disease recurrence risk in offspring of patient with ToF, anomalous pulmonary venous connection, abnormal connection and atrioventricular septal defect (AVSD) as 3.1%, 3.7%, 5.1% and 7.8% respectively, indicating very substantial genetic contributions to risk of these individually rare conditions. It is important to note here that "abnormal connection" refers to a category created for the study to include "abnormalities of situs, of atrioventricular concordance, and of the architecture of the atrioventricular and ventriculoarterial valves" (Burn *et al.*, 1998). None of the offspring of patients with TGA had been diagnosed with the disease. Using the disease recurrence risks for different categories of relatives, Burn *et al.* (1998) attributed different models of inheritance to the CHD subtypes analysed when possible. The variability of disease recurrence between CHD subtypes raises a number of issues. As in many other studies of the recurrence of CHD, the study by Burn *et al.* (1998) relies on relatively small sample sizes (Oyen *et al.*, 2009). The overall CHD recurrence will be influenced by the subtypes of CHD that are included in the study given the variability that exists between subtypes. It also raises the issue of how one might partition CHD into categories so as to provide the best estimate of recurrence risk in offspring while including enough data in each category for the resulting estimate to be meaningful. The exhaustive Danish Civil Registration System (DCRS) allowed Øyen *et al*. (2009) to avoid many of these problems. The study focused on information stored by the DCRS on 18,708 individuals born with CHD between 1977 and 2005 out of a total of 1,763,591 individuals. From the population data, Øyen *et al*. (2009) were able to calculate recurrence risk ratios for different subtypes of CHD. The recurrence risk ratio compares the recurrence risk of a disease trait in relatives of the probands with the share of the population diagnosed with the disease. Øyen *et al*. (2009) found strong variations in the recurrence risk ratio among different subtypes of CHD, "ranging from 3-fold to 80-fold increases compared with the population prevalence". This once more clearly demonstrates the existence of a strong genetic component to CHD while also emphasizing the variability in this component's importance depending on CHD subtype. Reproduced in Table 1 are the recurrence risks for first-degree relatives associated with each CHD subtype as they appear in Øyen *et al*. (2009).

CHD phenotype in first degree relative	**Relative risk**	**95% confidence interval**
Heterotaxia	79.1	32.9–190
Conotruncal defect	11.7	8.01–17.0
AVSD	24.3	12.2–48.7
Anomalous pulmonary venous return	N/A	N/A
Left ventricular outflow tract obstruction	12.9	7.48–22.2
Right ventricular outflow tract obstruction	48.6	27.5–85.6
Isolated septal defect	3.41	2.69–4.32
|_ ASD	7.07	4.51–11.1
|_ VSD	3.41	2.20–5.29
|_ ASD+VSD	5.57	0.79–39.5
Isolated PDA	8.74	5.58–13.7
|_ At term	4.8	0.68–34.1
|_ Preterm	19.5	10.5–36.1
Unspecified CHD only	5.22	3.40–8.00
Other specified CHD	12.6	7.68–20.5
Overall same heart defect phenotype	8.15	6.95–9.55

Table 1. Relative risk of recurrence of CHD by family CHD history among first degree relatives in Danish cohort. Table is reproduced from Øyen *et al*. (2009).

In chapter 4 and 5, I will explore what these studies said about recurrence risk for first-degree relatives of TGA and ToF patients respectively.

The knowledge surrounding environmental factors conducive to CHD remains limited. As CHD results from developmental events in the early stages of pregnancy, the focus of much of the research into environmental factors has been on maternal exposures (Gorini *et al.*, 2014). Gorini *et al*. (2014) note that maternal exposure to a range of chemicals with teratogenic properties, particularly in the first trimester of pregnancy, has been implicated in the development of CHD in the newborn. Excessive maternal stress or deleterious behaviours during pregnancy; such as smoking, can likewise increase the risk of a child being born with CHD (Jenkins *et al.*, 2007). Maternal chronic disease can also act as an environmental factor promoting CHD, as illustrated by a recent cohort study by Liu *et al*. (2013) carried out on upward of 2 million mother-infant pairs in Canada. The study uncovered multiple significant associations between child CHD and maternal chronic diseases such as hypertension and systemic connective tissue disorders. Excluding maternal CHD, maternal diabetes shows the strongest associations with CHD in infants, with reported odds of 4.65 and 4.12 of observing CHD in the offspring of mothers with type 1 and type 2 diabetes. Maternal diabetes has been associated with many CHD subtypes, including TGA, VSD, AVSD, PDA, hypoplastic left heart syndrome, cardiomyopathy and several conotruncal and outflow tract defects (Jenkins *et al.*, 2007). Blue *et al*. (2012) state that these strong environmental factors, along with strong genetic factors, account for 20% of CHD cases; the remaining 80% having some unknown multifactorial etiology.

The first CHD for which genetic factors were identified were those occurring as part of a syndrome; where the disease is accompanied by several other extracardiac defects (Bruneau, 2008). CHD is a common feature of trisomies 13, 18 and 21 as well as other chromosomal aneuploidies (Richards and Garg, 2010). CHDs are also identified as part of syndromes caused by copy number variants (CNV). For examples, CHDs are observed in a majority of cases of 22q11.2 microdeletion syndrome and Williams-Buren syndrome (Richards and Garg, 2010). Syndromic cases of CHD have also been found to be the result of SNVs in specific genes. This is the case for Holt-Oram syndrome which is triggered by mutations in *TBX5*, or Allagille syndrome which is caused by *JAG1*; or *NOTCH2* in some rare cases, both of which involve cardiac defects (Bruneau, 2008). Families in which CHD follows a Mendelian pattern of inheritance have also helped uncover genes that, when mutated, can cause isolated CHD. These include genes that play a role in early heart development such as *NKX2.5* and *GATA4* (Bruneau, 2008). However, with a majority of isolated CHD cases being sporadic, the specific contribution of genetic factors to CHD has been the source of much debate and mirrors to some extent that of other complex diseases. Of particular relevance are the developments that have taken place around common disease. Over the past few decades, two competing ideas have been advanced to explain common complex disease: The common disease, common variant (CDCV) and the common disease, rare variant (CDRV) hypotheses (Schork *et al.*, 2009). The CDCV hypothesis postulates that complex disease could be explained as the combined action of several common variants each with low penetrance (Schork *et al.*, 2009). In this context, 'common' is understood as an allelic population frequency greater than 1% (Manolio *et al.*, 2009). The CDRV hypothesis on the other hand predicts the existence of many highly penetrant rare variants, a single rare variant being sufficient to trigger complex disease (Schork *et al.*, 2009). Under this hypothesis, variants that contribute to complex disease are subject to negative selection and are therefore expected to be present at low frequency in the human population (Nelson *et al.*, 2012). A gene can be associated to a common disease phenotype through various rare variants acting within a specific region of the gene (Iyengar and Elston, 2007). While genome-wide association studies (GWAS) have identified thousands of common genetic variants that confer some of the risk in developing complex diseases, these variants only explain a modest fraction of inherited risk (Zuk *et al.*, 2014). This holds true for CHD as well. Looking at 1,995 CHD cases and 5,159 controls, a GWAS lead by Cordell *et al*. (2013a) found an association between ASD and the locus 4p16. However, this association only accounted for 9% of population-attributable risk for ASD specifically. As GWAS are built around the assumption that common complex disease follows the CDCV model, the 'missing heritability' has been used as an argument in favour of the CDRV hypothesis (Manolio *et al.*, 2009). The term 'missing heritability' here refers to the observation that many GWAS of complex disease only appear to explain a fraction of the estimated heritability for a given population; where heritability refers to the amount of trait variation that is due to genetic variation in a given population (Visscher *et al.*, 2008; Manolio *et al.*, 2009). The CDRV hypothesis has also gained traction as a result of population sequencing studies and what these revealed about the distribution of rare variants in the human population. The sequencing of 1,092 individuals by the 1000 Genomes Project Consortium (2012) has brought to light an abundance of rare variants in all human populations, exceeding what would be predicted from population genetic theory (Altshuler *et al.*, 2010). In this context, a variant is considered rare if it has a minor allele frequency (MAF) <0.5%. A study by Tennessen *et al*. (2012) of 15,585 protein coding genes in a large sample of individuals of European and African ancestry revealed that out of 500,000 identified single nucleotide variants (SNVs), 430,000 had a MAF<0.5%. The recent explosion of the human population over the last 10,000 years is partly responsible for the current population-wide distribution of rare variants (Keinan and Clark, 2012). Based on data from the 1000 Genomes Projects, Marth *et al*. (2011) showed that variants with a MAF<1% include a high proportion of functional variants. Comparing synonymous, missense and nonsense variants, Marth *et al*. (2011) found that the distribution of allele frequencies in these three functional categories of variants varied significantly across categories. A higher proportion of missense variants were found to be rare compared with synonymous variants while a higher proportion of nonsense variants were found to be rare compared to both synonymous and missense variants. Marth *et al*. (2011) reported that 78% of nonsense variants were found to have a MAF<1%. Variants that are predicted to be deleterious based on evolutionary conservation and potential protein structure or dosage changes are also overrepresented among rare variants (Marth *et al.*, 2011; Nelson *et al.*, 2012). Despite the current interest in the CDRV hypothesis, studies built around it have had limited success in identifying the underlying rare variants of many common diseases (Auer and Lettre, 2015). The study designs used for rare variant association and their limits are further explored in chapter 5.

As population studies such as those of Burn *et al*. (1998) and Øyen *et al*. (2009) make clear, CHD comprises a range of distinct subtypes governed by different levels of inherited risk. Importantly, estimates of the genetic component of any disease are influenced by factors such as the early mortality rate; including fetal loss, associated with the disease. Mortality rates vary substantially depending on the cardiac defect (Hoffman *et al.*, 2004). Severe cardiac defects are associated with high mortality rates early in life. For example, untreated ToF carries an 70% mortality rate in the first 10 years of life, TGA a 95% death rate in the first year of life (Starr, 2010; Saremi, 2014). Any mutation that increases the risk of developing severe CHD would therefore be expected to be strongly negatively selected (Cordell *et al.*, 2013b). Disease-causing variants in patients with severe CHD would be expected to be extremely rare or private, having arisen from a recent mutation event. In patients with no family history of CHD, the manifestation of

CHD can be the result of a *de novo* mutation occurring in a parental germ cell and being passed on to the patient. In chapter 4 and 5 I will describe how these observations influenced the approach to each CHD studied.

In a recent paper on the molecular genetics of CHD, Andersen *et al*. (2014) list a total of 55 genes associated with CHD identified via linkage analysis, the mapping of CNVs and WES. Using linkage analysis, a technique applied to large pedigrees that draws on recombination events to identify which of a set of genetic markers is in close proximity to a disease locus, Basson *et al*. (1997) were able to implicate the transcription factor gene *TBX5* in the development of Holt-Oram syndrome, an inheritable condition that can include heart defects such as ASD, VSD and cardiac conduction defects. Linkage analysis was also useful in elucidating disease-causing genes in isolated cases of familial CHD such as *NKX2.5* and *GATA4*, both also transcription factor genes (Schott *et al.*, 1998; Garg *et al.*, 2003). Three mutations in the transcription factor gene, *NKX2.5*, were uncovered by analysing four families with a history of CHD (Schott *et al.*, 1998). A single mutation in *GATA4* was found to be responsible for the various CHD phenotypes; systematically including ASD, established in one family over five generations (Garg *et al.*, 2003). In addition to transcription factors, the list of CHD associated genes presented by Andersen *et al*. (2014) includes genes involved in the signalling pathways of the heart, the cardiac sarcomere and chromatin modifiers. The authors also point to 500 genes that lead to cardiac defects in mice when mutated, inferring that there exists a similar number of disease genes for CHD in humans. Focusing on severe CHD, Zaidi *et al*. (2013) sequenced 626 parents-offspring trios, including 362 patients; the offspring, with serious cardiac defects not seen in parents. The parents-offspring trios were used to identify potentially disease-causing *de novo* variants occurring in human orthologs of genes highly expressed during cardiogenesis in mice. Based on the number of nonsynonymous *de novo* variants occurring in these genes in cases and controls, the authors conclude that *de novo* SNVs altogether contribute to approximately 10% of severe CHD, a contribution which they estimate rests on a set of 401 disease genes. Zaidi *et al*. (2013) also found an excess of *de novo* variants in patients with CHD in genes that affect histone methylation, potentially around key developmental genes.

1.3. Whole-Exome Sequencing (WES) Studies Of Disease

1.3.1. Sequencing technologies

The number of sequencing platforms has expanded dramatically in the past decade. In 2003, when the Human Genome Project (HGP) was completed, Sanger sequencing was the standard technique (Collins *et al.*, 2004). This technique was devised by Fred Sanger and his team as far back as 1977 (Sanger *et al.*, 1977). However, over the first decade of the millennium, a new range of massively-parallel sequencing platforms were developed and collectively referred to as next-generation sequencing (NGS) (Metzker, 2010). These platforms were developed to address a demand for fast, cheap and accurate high-throughput sequencing (Metzker, 2010). Many of these NGS technologies achieved this through the parallel sequencing by synthesis of distinct DNA fragments (Mardis, 2013). With the incorporation of each nucleotide during sequencing by synthesis, a chemical reaction emits a signal, indicating which nucleotide was added (Mardis, 2013). For this signal to be captured by the sequencing platform however, the DNA fragments must first have been amplified. Amplification is achieved through polymerase chain reaction (PCR) which creates local clusters of DNA fragments which signal synchronously the addition of each new base (Mardis, 2013). Figure 6 provides an example of the sequencing process on the Illumina platform.

The first commercial NGS platform, released in 2005, was the 454 pyrosequencing system. When first released, a single run on the 454 pyrosequencing platform was capable of delivering up to 25 million base pairs (Mbps) of sequence over four hours (Margulies *et al.*, 2005). This provided a net improvement on Sanger methods which, even today, reaches no more than 96 kbps per run (Shokralla *et al.*, 2014). Other NGS technologies include the Illumina platform, originally released in 2006, and the ABI SOLiD system, released in the following year. Each one of these platforms has been refined over the years. The latest versions of both these platforms, the Illumina HiSeq and the SOLiDv4, possess throughputs upward of a billion base pairs in their last iterations, as shown in Table 2 (Buermans and den Dunnen, 2014; Shokralla *et al.*, 2014). The research presented in this thesis revolves almost exclusively on sequence data produced via Illumina sequencing, one of the more widely used sequencers today. An overview of Illumina sequencing is presented in Figure 6. All three platforms rely on a high-definition camera to capture the signals emitted during sequencing by synthesis (Mardis, 2013). The Ion-Torrent, another massively-parallel sequencing platform, presents an innovation in that respect (Mardis, 2013). The Ion-Torrent captures a direct by-product of nucleotide incorporation, the release of a hydrogen ion, using an electric sensor (Mardis, 2013).

Sequencing platform	ABI 3730xl (Sanger)	454 GS FLX	Illumina HiSeq2500	ABI SOLiDv4
Output (per run)	96 kbps	450-700 Mbps	200 Gbps	300 Gbps
Reads (per run)	96	0.7-1 million	3 billion	2.7 billion
Read length (max)	1-1.5 kbps	700 bps	2 x 100 bps	85 bps
Run type (SE/PE)	SE	SE	SE & PE	SE & PE

Table 2. Current sequence platform specifications for Sanger sequencing and three NGS platforms. Numbers are from Shockralla *et al.* (2014) and Buermans and den Dunnen (2014). Output and read length are given in read base pairs. SE. single end reads; PE, paired end reads.

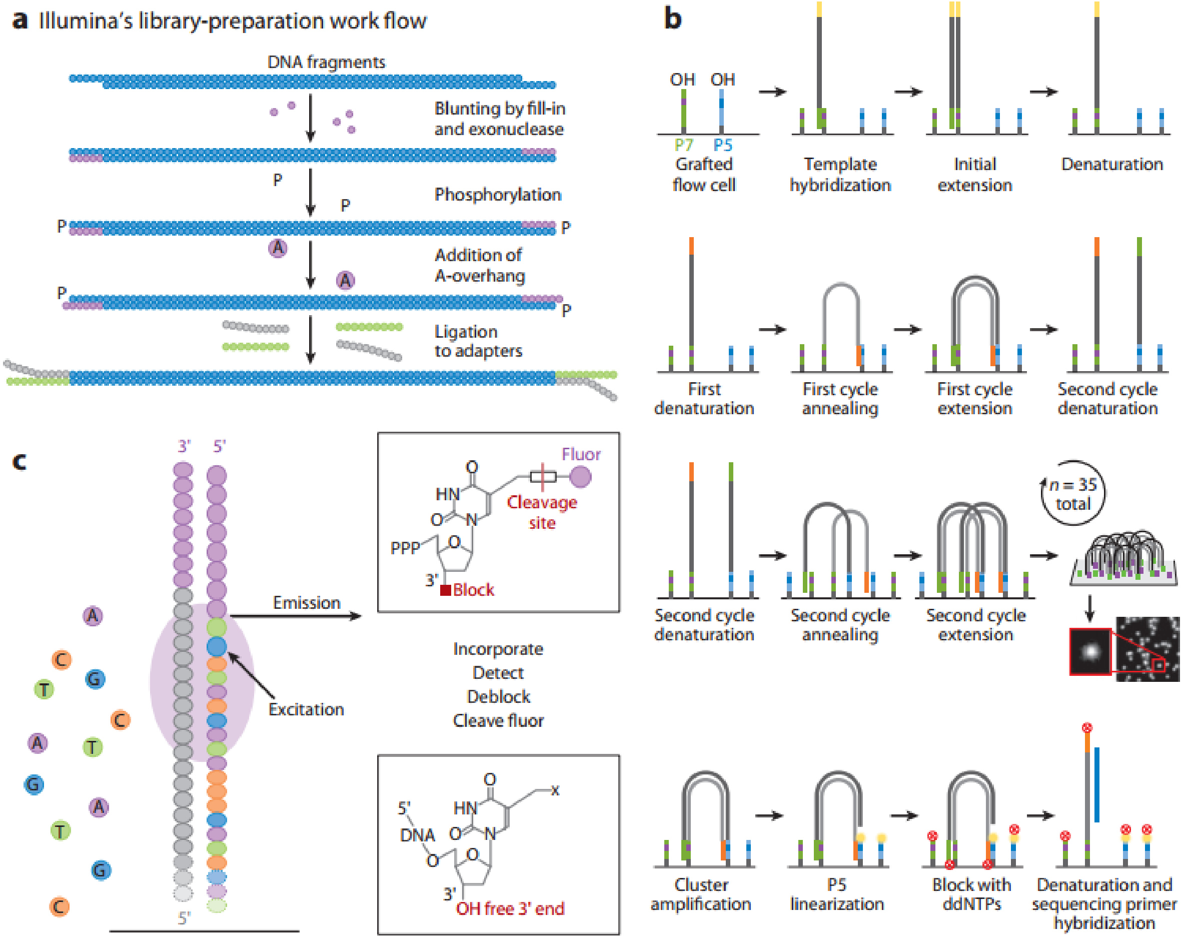

Figure 6. An overview of The Illumina sequencing method. Figure is from Mardis (2013). During library preparation (a) DNA fragments are flanked with adapter sequences. These adapters complement adapter sequences distributed across a flow cell. After hybridisation, the DNA fragments are amplified using bridge PCR (b). This generates local clusters of identical DNA fragments which give off a synchronous fluorescent signal during sequencing by synthesis (c).

Building on the successes of NGS, new technologies have emerged (Schadt *et al.*, 2011) . Unlike NGS technologies, these new technologies do not require an amplification step. A sequencer such as the Helicos Genetic Analysis platform is capable of detecting the signal emitted for a single DNA molecule (Schadt *et al.*, 2011). Some of these technologies present an additional innovation: sequencing in real-time (Mardis, 2013). The combination of these innovations is epitomised by the Pacific Biosciences' single molecule, real time (SMRT) system (Mardis, 2013). The SMRT system is able to complete an entire run within hours where the original NGS technologies take days (van Dijk *et al.*, 2014). Another advantage with many of these new technologies is the ability to sequence reads orders of magnitude longer than what can be achieved with NGS technologies (Koren and Phillippy, 2015). The SMRT system can sequence reads 50 kbps (Koren and Phillippy, 2015). A more recent, technological innovation, the Oxford nanopore minION, does not have a well-defined upper limit; its maximum read length dependent on the length of the DNA strand that passes through it (Karlsson *et al.*, 2015). Platforms such as Oxford nanopore platform no longer use sequencing by synthesis, using instead a molecular structure called a nanopore capable of detecting nucleotides passing through it (Koren and Phillippy, 2015). One important thing to note however is that many of these new technologies do not yet provide the read accuracy necessary to investigate SNVs and small indels (Koren and Phillippy, 2015). These technologies are nonetheless useful for other purposes such as finishing genome assemblies previously hindered by short reads (Koren and Phillippy, 2015).

1.3.2. Exome capture via target-enrichment

NGS technologies have greatly enhanced the search for variants underlying various disease traits. By sequencing twelve unrelated individuals, four of which were affected by Freeman-Sheldon syndrome, a rare autosomal dominant disorder, Ng *et al*., (2009) first illustrated the potential of such practice. After filtering variants found in the four Freeman-Sheldon syndrome patients against those found in eight HapMap exomes and a database of known variants, dbSNP (Sherry *et al.*, 2001), Ng *et al*., (2009) found that only a single gene, *MYH3*, had corresponding nonsynonymous rare variants in all four patients. Dozens of studies have since identified the genetic underpinnings of many disorders, including extremely rare conditions such as Kabuki syndrome (Gonzaga-Jauregui *et al.*, 2012). What many of these studies have in common is that they favour the use of whole-exome sequencing (WES) over whole-genome sequencing (WGS). Examples of WGS being used to successfully identify disease variants do exist, as illustrated by a study conducted by Nishiguchi *et al*. (2013) which established not only new disease-causing variants, but also a previously unidentified disease gene for a progressive blindness disease, retinitis pigmentosa. Furthermore, WGS has come to play an increasingly larger role in clinical practice, a fact exemplified by large-scale initiatives such as England’s 100,000 genomes project (Siva, 2015) which includes patients with a whole-range of rare diseases and cancers and their relatives. However, the costs involved in both sequencing and storage have long been prohibitive for that line of research (Rabbani *et al.*, 2014). Teer and Mullikin (2010) originally referred to WES as the “sweet spot before [WGS]”. WES offers a relatively cheaper alternative to WGS by focusing the sequencing effort on exonic regions of the genome (Teer and Mullikin, 2010). In humans, the exome accounts for only 1% of the entire genome, but contains all of the genome’s protein-coding sequences (Teer and Mullikin, 2010). Protein-coding regions alone are estimated to harbour 85% of deleterious mutations in the whole genome (Majewski *et al.*, 2011).

In order to focus sequencing on exome regions, an exome capture step is required to complete sample processing. The library of genomic DNA fragments obtained for sequencing is subjected to target enrichment, a process by which specific DNA fragments are captured using a custom set of complementary oligonucleotides which correspond to a selected fraction of the genome (Mamanova *et al.*, 2010). While various capture methods exist, hybrid capture is the most prominent for whole-exome capture. The most commonly used exome capture kits; offered by Nimblegen, Agilent and Illumina, use in-solution hybrid capture (Chilamakuri *et al.*, 2014).

Hybrid capture for NGS was first developed by Nimblegen with on-array capture. In a paper introducing their approach, Hodges *et al*. (2007) describe the repurposing of DNA microarray for exome capture and sequencing. The DNA microarray hybridisation step is used to capture exon targets. The next step for a standard DNA microarray experiment would consist in scanning the microarray to detect some chemically-induced signal given off by labelled targets (Mamanova *et al.*, 2010). However, in exome capture, the targets are not labelled and are instead recuperated for sequencing (Mamanova *et al.*, 2010).

Building on the success of on-array capture while addressing some of its drawbacks, in-solution capture hybridisation was developed (Mamanova *et al.*, 2010). With in-solution capture, probes are eluted from the DNA microarray and amplified through PCR and subsequently biotinylated. The amplification step creates an excess of probes, which in turn allows for target-enrichment to take place with less target DNA. The fragments hybrised in a solution are then captured by beads that bind biotin (Gnirke *et al.*, 2009). Recent hybridisation techniques use in-solution capture. Table 3 lists the principal exome capture technologies.

Exome capture technologies	**Agilent SureSelect Human All Exon (v.4)**	**Roche NimbleGen SeqCap EZ Exome (v.3)**	**Illumina Nextera & Truseq Exome Enrichment***
Targeted bases	51.1 Mb	64.1 Mb	62.08 Mb
Number of targets	185,636	368,146	201,071
Overlap with…			
RefSeq	28.2 Mb	29 Mb	31.9 Mb
CCDS	27.4 Mb	28.02 Mb	31.02 Mb
ENSEMBL	29.43 Mb	30.15 Mb	32.07 Mb

Table 3. Target specifications of the four principal exome capture technologies and their overlap with various databases reporting exons. Each technology covers different regions of the human exome. Only 26.2 million bases (Mb) targeted are shared by all four capture technologies. The data is obtained from Chilamakuri *et al*. (2014). *Illumina's Nextera and Truseq Exome Enrichment technologies use the same targets.

1.3.3. Exome-sequencing study tools

In order to extract useful information from sequence data originating from WES, a number of steps are required. The first step of any sequencing analysis is a thorough quality control of the data (Altmann *et al.*, 2012). Quality control is followed by the alignment of sequence data to one or more reference sequences and the detection of potential variants. Depending on the aim of a WES analysis, the variant data can be annotated and filtered according to a set of specific criteria. An overview of the WES pipeline is provided in Figure 7.

For the majority of quality control, alignment and variant calling tools, an estimate of base quality is essential (Altmann *et al.*, 2012). NGS platforms are equipped with base calling software which determines the quality of the data captured by the sequencer (Ledergerber and Dessimoz, 2011). The signals emitted during sequencing-by-synthesis and recorded by an image-capture device, are transcribed as base calls (Altmann *et al.*, 2012). Each base call assignment is accompanied with a value indicating the certainty with which the base call was made. This value takes into account properties of the signal such as intensity or distance from other clusters, as well as sequencing cycle (Altmann *et* *al.*, 2012). This base call quality is stored as a Phred-like score (Ewing and Green, 1998; Ewing *et al.*, 1998; Altmann *et al.*, 2012). The base call quality is the first assessment of sequencing quality and is crucial for quality control. A number of recent software programs attempt to improve base quality score assignment, either by providing a recalibration of the calling (DePristo *et al.*, 2011) or offering a different base calling procedure to the one provided with the platform (Massingham and Goldman, 2012). Despite the latter developments, the sequencer's own base quality scoring system remains the most widely used (Altmann *et al.*, 2012).

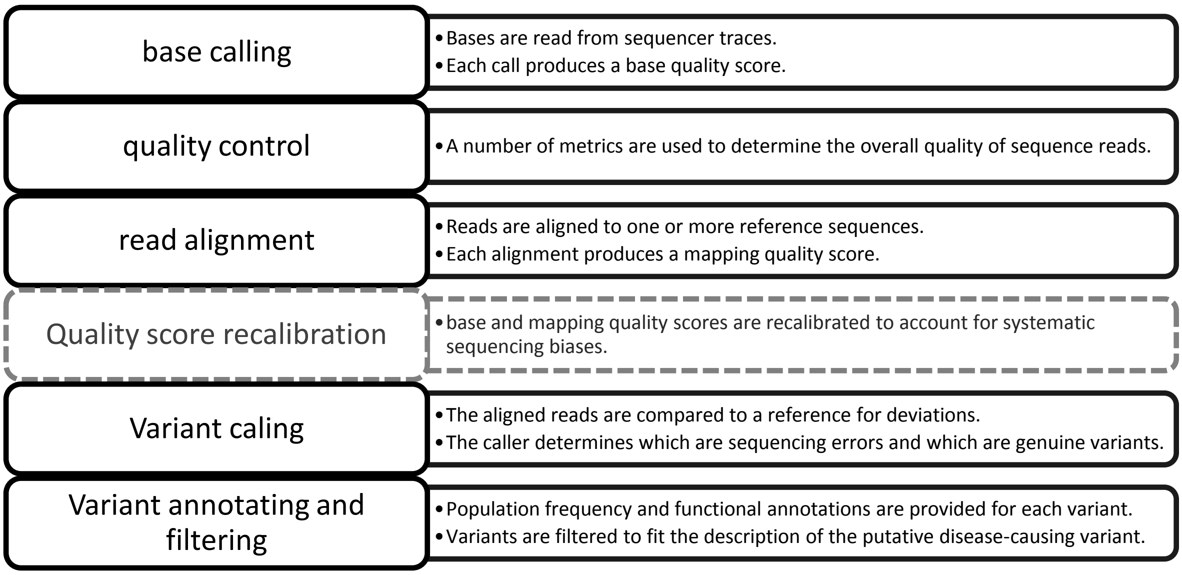
base calling
• Bases are read from sequencer traces.
• Each call produces a base quality score.
quality control
• A number of metrics are used to determine the overall quality of sequence reads.
read alignment
• Reads are aligned to one or more reference sequences.
• Each alignment produces a mapping quality score.
Quality score recalibration
• base and mapping quality scores are recalibrated to account for systematic sequencing biases.
Variant caling
• The aligned reads are compared to a reference for deviations.
• The caller determines which are sequencing errors and which are genuine variants.
Variant annotating and filtering
• Population frequency and functional annotations are provided for each variant.
• Variants are filtered to fit the description of the putative disease-causing variant.

Figure 7. A typical WES workflow. Some variant callers such as SOAPsnp or GATK require a quality score recalibration step.

Some measure of quality control is important at every stage of sequence analysis. However, a clear emphasis is placed on the quality control of raw sequence data as problems arising during or prior to sequencing are likely to compromise the entire analysis (Guo *et al.*, 2013). At several steps during sample processing and sequencing, data might become significantly compromised. For example, during sample and library preparation, DNA can undergo degradation or contamination (Robasky *et al.*, 2014). PCR amplification can be marred by excessive amplification errors if the genomic data is of poor quality (Robasky *et al.*, 2014). During sequencing, signals from DNA fragment clusters might overlap, usually as a result of overloading of the sequencer's flow cell (Robasky *et al.*, 2014). Quality control provides the means to assess the integrity of sequencing data before any analysis takes place. Many of these problems can be detected by studying the distribution of values for some metric in the pool of raw sequence data (Ross *et al.*, 2013). This usually includes metrics such as per base sequence quality and content, per sequence mean quality score and content, sequence length (Patel and Jain, 2012). Single value metrics, such as the estimated rate of duplications in the data, can also help flag up problems with sequencing (Patel and Jain, 2012). In principle, any indication of overall poor quality data should result in a new round of sequencing once the source of error has been addressed. Another important stage requiring quality control is variant calling. Estimating the sensitivity and specificity with which calls were made can help assess the reliability of results at every stage. In this context, sensitivity refers to the rate with which a caller identifies variants that are present in an individual while specificity refers to the rate with which callers exclude variants that are not present in an individual. Microarray data has often been used for the purpose of estimating variant calling sensitivity and specificity (Houniet *et al.*, 2015). However, microarray data is not always available. Publicly available population data can also provide a reliable means of estimating sensitivity and specificity (Houniet *et al.*, 2015). Some quality control at each stage is important in order to correctly interpret WES study results.

Even when focused on the exome, sequencing produces a large amount of read data. Coupled with the relatively short size of each read and the presence of base miscalls, determining the original position of each read in the genome proves to be a computer-intensive task (Altmann *et al.*, 2012). This is rendered all the more complex by regions of the genome that contain repeats or that are homologous to other regions of the genome. Sequence alignment is generally achieved by mapping particular sequence reads to regions of the genome by comparing these to one or more reference sequences. In the case of human genome, reference sequences have two principal sources: the university of California Santa Cruz (UCSC) genome assembly and the Genome Reference Consortium (GRC) genome assembly (Pabinger *et al.*, 2014). Both assemblies are identical, differing only in presentation (Pabinger *et al.*, 2014). In order to provide fast alignment of large amounts of short read data, various data structures are used to efficiently access subsets of the reference data, read data or both (Li and Homer, 2010). Fast aligners can be broadly divided into two categories: those with algorithms based on hash tables and those based on prefix/suffix tries (Li and Homer, 2010). Hash tables are data structures that link objects, referred to as keys, to corresponding values in a table. The key provides the location of the corresponding value through a transformation by a hash function. In this context, hash tables can either be used to store information on k-mer subsequences of the reference; a solution employed by NovoAlign (Novocraft, 2014b), or information on the reads themselves, an approach used by MAQ (Li *et al.*, 2008). In either case, a particular DNA sequence acts as a key for a list of values representing the locations at which the particular combination occurs. A prefix/suffix trie; derived from the word "retrieval", is an ordered tree data structures from which a value can be retrieved using the suffixes; or prefixes, of a sequence of characters as a key. In the context of alignment, the values will refer to the position at which a subsequence starts; or ends if prefix trie. An example of a prefix trie is given in Figure 8.

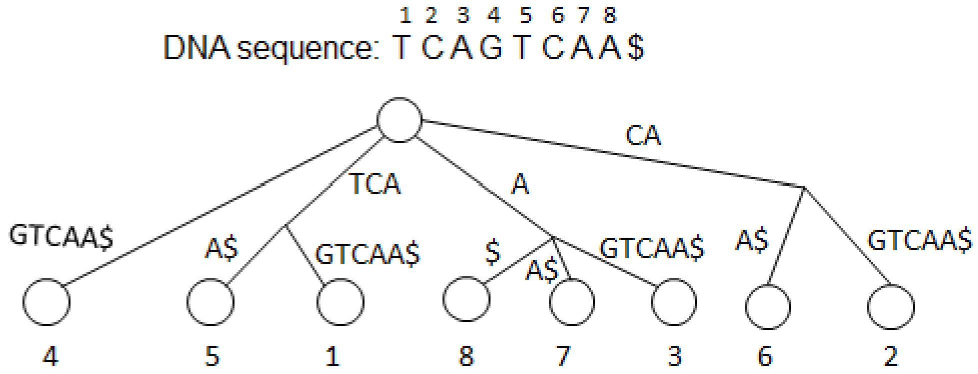

Figure 8. Example of a DNA sequence stored in a suffix tree. Each possible suffix subsequence can be used as a key to retrieve the index at which the subsequence starts. Each edge corresponds to a subsequence starting with a new character (for the root node: A, T, G and C). For example, TCA can either be followed by A or G, leading to two edges.

Prefix/suffix tries can take on one of many forms such as: suffix-tree, enhanced suffix array and FM-index (Li and Homer, 2010). The widely-adopted aligners in this category such as Bowtie (Langmead *et al.*, 2009) and BWA (Li and Durbin, 2009) employ FM-index (Li and Homer, 2010). As with base calling, the quality of each read mapping is given as a Phred score (Li and Durbin, 2009). In this case, the score represents the confidence with which a read is aligned. This depends on the number of mismatches between a read and its best alignment to the reference (Li and Homer, 2010). In addition to base mismatches, most aligners also accept alignments that leave gaps in either the read or the reference. When these types of gaps are not errors, they are indicative of the presence of a small insertion or deletion. Together, the base and mapping quality scores are useful in determining the existence of deviations from the reference corresponding to SNVs, insertions and deletions.

Aligned sequence reads provide a data set from which SNVs and small indels can be uncovered. This process is referred to as variant calling (Nielsen *et al.*, 2011). Many methods complement variant detection with genotype prediction (Altmann *et al.*, 2012).

In this case, the whole process is referred to as genotype calling; variant calling being implied. Early implementations of variant calling relied on counting the occurrences of nucleotides mapping to a particular position of the reference sequence (Altmann *et al.*, 2012). Any deviation from the reference sequence that was observed in a set of high-quality reads above a certain frequency was called as a variant. While this was sufficient for revealing variants in loci combining high-quality reads with high read depth, the approach did little to address the uncertainties due to errors in base calling and alignment and low read depth (Altmann *et al.*, 2012). Whichever exome capture technology is used, there is inevitably some variability in sequence coverage; or read depth (Chilamakuri *et al.*, 2014). Additionally, high or low GC content reduces the efficiency of PCR amplification, leading to a low read depth in the corresponding regions (Aird *et al.*, 2011). This early approach has therefore been supplanted by a number of callers using probabilistic methods to distinguish genuine variants from sequencing errors (Pabinger *et al.*, 2014). With few exceptions, variant callers use the Phred base and mapping quality scores as input for their probabilistic models (Pabinger *et al.*, 2014). Methods built on Bayes' theorem, such as those employed by SAMtools (Li *et al.*, 2009a) and GATK (DePristo *et al.*, 2011), produce from the sequence data a posterior probability of an observed deviation being a variant.

The sequencing and variant calling of large number of samples has become common practice in WGS studies, particularly when applied to population and cancer genomics studies (Chen and Sun, 2013; Wang *et al.*, 2013). Originally, population studies employed pooled sequencing, where DNA from a cohort of patients was included in a single library preparation (Altmann *et al.*, 2011). This was done to reduce costs incurred by sample preparation of a large number of individuals (Altmann *et al.*, 2011). However, in recent years, population studies have returned to a more traditional approach. The study of ToF patients presented in chapter 5 is one such example. In cancer studies, sequencing is performed on DNA from tumour and normal cells in order to uncover somatic SNVs which may contribute to a particular type of cancer (Larson *et al.*, 2012). Identifying variants in these types of sequence data requires callers to work with a different set of assumption. Population and cancer studies therefore have their own range of specialised variant callers. In WES, multiple sample sequencing and concurrent variant calling can be used to detect a variant inherited by multiple individuals in a family. Using the family structure to draw relationships between the different samples, callers such as FamSeq

(Peng *et al.*, 2013), PolyMutt (Li *et al.*, 2012) and DeNovoGear (Ramu *et al.*, 2013) provide joint genotype predictions that are consistent with Mendelian inheritance. More details on these approaches to variant calling will be given in Chapter 3.

Variant calling produces a large list of variants. In order to pinpoint a subset of variants that match the description of the putative disease-causing variant, filtering and annotation is required (Altmann *et al.*, 2012). The first step is to filter for exome data. Variants that fall outside the exome target region are filtered out. Two forms of filtering are then applied following variant calling. The first aims to reduce the burden of false-positive variants in the list by subjecting it to quality thresholds. For example, variants supported by very low read-depths can be unreliable and are therefore often excluded (Nielsen *et al.*, 2011). The second form of filtering aims to reduce the list of variants to a set most relevant to the observed disease and its hypothesised pattern of inheritance (Altmann *et al.*, 2012). For example, an autosomal recessive disease in a patient with consanguineous parents will require that the individual manifesting the disease be homozygous for the variant of interest while parents carrying the disease are heterozygous. Variants present in individuals that are not expected to carry the disease-causing variant or absent in individuals that are expected to carry the disease-causing variant can likewise be filtered out. Additional data is often required for further filtering and variant prioritisation such as population frequency data and functional prediction (Wang *et al.*, 2010). Variant databases such as dbSNP (Sherry *et al.*, 2001) or the 1000genomes project (Abecasis *et al.*, 2012) report countless variants and their corresponding MAFs. If the studied disease is predicted to be the result of a very rare allele, variants that have a high MAF can be filtered out, leaving only rare variants; typically understood as variants with a MAF<1% (MacArthur *et al.*, 2014). Provided there are no obvious candidates among the remaining variants, these can be prioritised according to their putative functional impact on protein structure. A first distinction can be made between potentially truncating variants; nonsense variants, frame-shifting indels and splice site variants, and other non-synonymous variants. For missense variants, a number of functional predictions can be used to support a variant's deleterious potential (Dong *et al.*, 2015). These functional predictions are based on such things as sequence conservation at the variant site and the predicted effect of the variant on specific protein structure features (Dong *et al.*, 2015). The assumption is that the disruption of a highly-conserved sequence or of a crucial protein structure is likely to be damaging. However, many of these functional prediction tools were not originally intended as predictors of disease-causing potential (Grimm *et al.*, 2015). As a consequence of the different criteria underlying different functional predictors, the results for a single variant occasionally appear contradictory (Dong *et al.*, 2015). The accuracy of functional predictors is limited. Dong et *al*. (2015) measured the qualitative prediction performance of 11 functional predictors against a dataset of known deleterious and neutral variants and found MutationTaster to have the highest estimated accuracy with 86%. Another popular prediction tool, PolyPhen-2 had an estimated accuracy of 79%. One should therefore be careful not to over-interpret the predictions obtained from these programs. There is a number of services that integrate a variety of databases and functional annotation tools into a single pipeline, as is the case with ANNOVAR (Wang *et al.*, 2010).

As a result of additional filtering through annotation, variants can be concentrated into a small set of potentially disease-causing variant. The best candidates can be selected for sequence validation and functional studies aimed at confirming the link between mutation and disease trait.

1.3.4. Exome-sequencing study designs

WES studies aimed at identifying disease-causing mutations can take on a number of forms ranging from a simple family-based study to larger case-control studies (Bamshad *et al.*, 2011). What particular study design is most appropriate will often depend on the type of disease and mutation being studied and the availability of samples. Family-based studies are more adequate when the disease trait is identified as segregating in a family in a way that suggests the inheritance of a single causative variant. A large number of individuals with the disease might not be available for sequencing, but multiple individuals in a single family might exhibit the same disease phenotype. To date, the genetic underpinnings of dozens of rare syndromes have been identified through family-based NGS studies (Gonzaga-Jauregui *et al.*, 2012). On the other hand, given a large set of individuals with putative rare *de novo* mutations triggering a single disease, a large case-control study may permit the identification of small regions or genes in which variations occur more frequently in a case population compared with controls. Because of the cost involved in sequencing large numbers of case and controls, the latter type of study is still rare (Derkach *et al.*, 2014). This type of study will be discussed in more depth in chapter 5.

WES study designs are often inspired by previous methods of identifying disease-causing mutations which relied on different technologies. From the 1980s onwards, highly penetrant mutations were identified through linkage analysis (Bailey-Wilson and Wilson, 2011; Brunham and Hayden, 2013). Taking advantage of the recombination that occurs during meiosis, linkage analysis uses markers to identify regions in the genome that co-segregate with a disease phenotype in a family. The regions can then be subjected to targeted sequencing so that the precise molecular alterations can be identified. Some early examples of successful linkage analysis include the identification of a disease locus for Huntington's disease and cystic fibrosis (Gusella *et al.*, 1983; Riordan *et al.*, 1989). Brunham and Hayden (2013) report that between 1980 and 2010, the genetic basis of upwards of 4,000 Mendelian diseases was elucidated through linkage analysis, particularly after the sequencing of the human genome at the turn of the century. The sequencing of the human genome enhanced linkage analysis by providing investigators with better knowledge of the genes within a locus of interest. However, the introduction of NGS technologies has rendered linkage studies largely superfluous. There exist some examples in the literature of WGS and linkage analysis being used in tandem. One example is the discovery of deletion in the gene *PTPN11* in a family with metachodromatosis, a disease that affects bone development (Sobreira *et al.*, 2010). The region in which that deletion was identified was first determined through linkage analysis. A single individual from that family was then analysed using WGS and the deletion was uncovered (Sobreira *et al.*, 2010). Nonetheless, WES alone can be used to identify highly penetrant alleles and does not require as many samples or as many familial generations (Brunham and Hayden, 2013). This makes NGS the ideal technology for investigating rare diseases in families, even with small pedigrees, making linkage studies redundant.

The elucidation of diseases through linkage analysis was progressively supplanted by genome-wide association studies (GWAS) as the focus of research moved from Mendelian diseases to common complex disease (Antonarakis and Beckmann, 2006). Theoretically, GWAS provide a better approach for elucidating diseases that follow the common disease common variant (CDCV) model, deriving their strength from large numbers of unrelated case and control samples (Saint Pierre and Genin, 2014). GWAS use the case-control design in order to identify differences in allele frequency at known polymorphic sites. These polymorphic sites in turn provide an association between a particular locus containing the causal variant; in linkage disequilibrium with the polymorphic site, and a disease trait. While GWAS has successfully associated a number of common diseases with various loci, these loci often only account for a small fraction of the estimated heritability in the population under study (Manolio *et al.*, 2009). Efforts have been made to adapt GWAS to detect the effect of aggregates of rare variants (Bansal *et al.*, 2010). In order for the actual rare variants to be detected, GWAS need to be supplemented by NGS studies with which the loci producing the strongest associations can be analysed for individual variants (Nejentsev *et al.*, 2009). Variants that are extremely rare in the human population; MAF<0.5%, will be undetectable in GWAS however due to low power unless a very large number of samples are used. This warrants a new strategy. It is in this context that large WES case-control studies are expected to take over from GWAS (Cirulli and Goldstein, 2010). The main obstacle remains the cost of sequencing large numbers of cases and controls (Derkach *et al.*, 2014). The size of case and control sets required to identify rare variants is discussed at length in chapter 5.

In WES studies, closely related individuals can be used to uncover both inherited and *de novo* disease-causing variants provided the variants exhibit high penetrance. Studies of *de novo* variants, concentrate on individuals exhibiting a disease trait which is not observed in their parents and is suspected to be the result of *de novo* variants (Bamshad *et al.*, 2011). Here the assumption is made that the disease-causing variants are dominant and do not have incomplete penetrance (Bamshad *et al.*, 2011). A variant with incomplete penetrance could be inherited without leading to disease in all carriers. *De novo* studies could therefore be used to investigate traits that are highly deleterious and occur only in sporadic cases (Bamshad *et al.*, 2011). *De novo* variants are identified by sampling parents-offspring trios or quartets in which a child is alone in exhibiting a disease trait. Sequencing data from parents and sibling provide a set of variants which can be excluded from the child (Bamshad *et al.*, 2011). The assumption made here is that a variant occurring in two closely related individuals is unlikely to have occurred as the result of two distinct mutation events and thus more likely to have been inherited. Assuming perfect sequencing, the remaining list of variants would mostly contain *de novo* variants, with the remainder of variants undetected in parents due to mosaicism; where a particular variant is only present in certain cell types in an individual. This is the study design followed in chapter 4. Familial relationships can also be drawn on to identify variants underlying a disease trait. If a disease trait follows a Mendelian pattern of inheritance, it can be suggested that it is being caused by a single causal variant, provided this variant shows high-penetrance. Assuming high-penetrance is essential in determining carrier status and thus ultimately identifying a causal variant. Candidate variants are identified by sequencing related individuals; healthy and affected, and selecting those variants present in carriers and absent in non-carriers. Depending on the pattern of inheritance, further assumptions can be made about the location of the putative variant or the genotype expected in each carrier.

Case-control studies are largely focused on unrelated individuals and seek to identify causal regions or genes in the development of a disease. Case-control studies focused on sporadic disease for which *de novo* variants are suspected are the exception. Siblings are assumed not to carry disease-causing variants and can therefore be used as controls. This was the strategy adopted for a large autism spectrum disorder study by Sanders *et al*. (2012) in which 238 families were sequenced. These families consisted of 38 parents-offspring trios and 200 parents-offsprings quartets. The 200 non-affected siblings were used as controls, revealing an excess of non-synonymous *de novo* variants in probands and uncovering a gene association between autism and the gene *SCN2A*. A recent example of a case-control study for CHD is the study by Zaidi *et al*. (2013). This study investigates potential causal genes for an array of severe CHD phenotype. The study consists of 362 parents-offspring trios where the proband has severe CHD, and 264 parents-offspring controls. I use the case-control WES study design in chapter 5 and it will be discussed at greater length in that chapter. As the focus of WES case-control studies shifts towards rare variants of lower effect sizes, increasingly larger case-control studies will be required to increase statistical power (Lee *et al.*, 2014b).

1.4. Conclusion

With this PhD project, I investigate the genetic underpinnings of severe CHD using WES, with a particular emphasis on TGA and TOF. Based on current research on the genetic etiology of both these diseases, I conducted bioinformatics analyses of WES studies and identified genes that hold potentially disease-causing variants. Additionally, I devise and implement a new approach to variant calling with the aim of improving detection of variants shared by individuals carrying a disease when a multiply affected, putative Mendelian pedigree is being studied.

Chapter 2. Methods

2.1. Overview

Whole-exome sequencing plays a central role in the research presented in this thesis. In this chapter, I will cover the core methods of WES used in data generation. I will first specify which NGS and exome capture technologies were used. I will also review the recurring tools and metrics used to assist exome data analysis throughout the entire project and provide an outline of the hardware enlisted for the analysis. However, since bioinformatics methods and pipeline development suitable for the three different study designs adopted in my work; analysis of Mendelian families; analysis of *de novo* variants in trio families; and case-control analysis, formed a significant portion of my work, each individual chapter of this thesis has its own set of methods developed by me, complementing the standard methods described in this chapter.

2.1.1. Genome sequencing and exome capture

While the sample processing, sequencing and exome capture of genomic DNA was not performed by me personally, reviewing these aspects may help to understand the various analyses I performed. I will therefore briefly describe some of the steps involved in the generation of WES data.

2.1.2. Sequencing and capture

The required sequencing for this project; including preliminary steps such as sample processing and library preparation, was divided between three different research centres:

- The Institute of Genetic Medicine (Newcastle University, UK)
- Glasgow Polyomics (University of Glasgow, UK)
- The McGill University and Génome Québec Innovation Centre (Montreal, Canada)

DNA was extracted from blood samples originally held at Newcastle's Institute of Genetic Medicine (IGM). Target enrichment for exome capture was performed using SureSelect^{XT} Human All Exon 50Mb kit V4 (Agilent Technologies, 2015) from Agilent Technologies, USA. Details on the performance of this technology compare to other available technologies can be found in Chapter 1.

At all three locations, sequencing was performed using the Illumina platform (Illumina, 2015c). Both UK sites used the Illumina Genome Analyzer IIx (GAIIx) (Illumina, 2012b), while The McGill University and Génome Québec Innovation Centre (MUGQIC) used Illumina HiSeq 2000 sequencers (Illumina, 2014a). The Illumina GAIIx produces around 30 million reads per flowcell lane. Each read is 76 base pairs (bp) long. The Illumina HiSeq 2000 produces around 180 million bps per flowcell lane. The reads are 101 bp long. The HiSeq 2000 is able to process large numbers of samples together using multiplex sequencing. All sequencing was done with paired-end reads (Illumina, 2015b). Over the course of three years, the sequencing shifted from the IGM to Glasgow Polyomics (GP) and finally to the McGill University and Génome Québec Innovation Centre (MUGQIC).

2.1.3. Quality control

Before proceeding with sequence analysis, the quality of the raw exome data must be assessed. This preliminary quality control step was carried out at the sites where the sequencing was performed and some samples may have been sequenced a number of times as a result. The quality metrics used in the MUGQIC to assess sequencing quality were provided together with the sequencing data. For data emanating from UK locations, I have independently produced quality metrics using the NGS QC toolkit v2.2.3 (Patel and Jain, 2012). In either case, the quality metrics, using the platform's base quality score (Ewing and Green, 1998; Ewing *et al.*, 1998), included the following:

- By base position in the read
    - Average base quality score
    - Percentage of reads with base falling within a base quality score range (i.e. 0-10, 11-20, 21-30, 31-40)
- By number of reads
    - Read falling within a GC content percentage bracket. (i.e. 0-5,…,95-100)
    - Read average base quality score
- Percentage base composition (including non-ATGC)

The NGS QC toolkit also provided a summary of its quality checks, giving a percentage of high and low quality reads. By default, NGS QC toolkit sets the threshold of high-quality at 20 which corresponds to a base calling accuracy of 99%. I also applied this threshold to quality metrics generated by MUGQIC. Using these quality metrics, I was able to check for irregularities which would betray poor quality data. The first step was to check that high-quality reads represented ≥90% of the data. Following this, I reviewed metrics that presented base quality scores by base position in the read to check for any sudden drops in quality. I verified that there was no excess of non-ATGC bases and that none of the four bases were either significantly over or underrepresented. In addition to the previously described quality metrics, MUGQIC also provided the percentage of read duplicates; PCR and optical duplicates, per sample providing some additional insight into the quality of sequencing at MUGQIC.

2.2. Data analysis

2.2.1. Hardware specifications

The bulk of data analysis was completed on a computer cluster based at the IGM, currently running on a Scientific Linux 6.3 (Fermilab, 2014) operating system. Jobs were submitted to the cluster using the OGS/GE 2011.11p1 batch-queuing system (Open Grid Scheduler project, 2013) for distributed resource management. The hardware on which the operating system runs has been upgraded on multiple occasions over the past three years.

At the start of my PhD project, the cluster was composed of a single headnode and 16 compute nodes. A Dell R510 server node was used as the headnode, while the 16 compute nodes were distributed across 4 Dell C6100 rack servers. The headnode operates on two Intel Xeon E5620 2.40 GHz quad-core processors and 12GB of RAM and has 19TB of available storage space. Each of the compute nodes operates on two Intel Xeon E5640 2.67 GHz quad-core processors, 47GBs of RAM and has 160GB of hard disk space.

In the fall of 2012, a login node and 4 additional compute nodes were added. The additional compute nodes share the same specifications as the previous 16 with the exception of RAM; 96GBs, and hard disk space; 900GBs. A Dell C1100 server node acts as the login node with two Intel Xeon E5640 2.40 GHz quad-core and 24GBs of RAM. In 2014, a high-intensity compute node; with four Intel Xeon E7-4820 2.00 GHz 8-core processors, 520GBs of RAM and 400GBs of hard disk space, was added to the cluster. Such upgrades were crucial in order to pursue large scale WES analyses. In addition to new login and compute nodes, 2 Dell R510 nodes attached to 2 Dell MD1200 storage shelves provide 73TB of lustre storage.

In the final year of the PhD project, a computer cluster at Manchester University's Faculty of Medical and Human Sciences was also made available for this project. The cluster was primarily used for the data analysis of 871 ToF cases and 500 controls as presented in chapter 5. The cluster comprises 10 compute nodes; each equipped with 512 GBs of RAM and a dual 2.60 GHz 8-core processor. The compute nodes provide a total of 55TB of hard disk space.

2.2.2. Software and scripts

WES analysis calls upon a whole sequence of programs, requiring some degree of automation. Analysis pipelines were devised through a combination of Perl (Perl.org, 2002) and Bash shell (GNU Project, 2014) scripts which could then be submitted as one or more jobs on a cluster's compute nodes. The scripts were also used to directly filter and interpret data without the use of any third-party software. The scripts used for this project thesis are a combination of my modification of scripts originally written by colleagues and scripts written by myself to fulfil study-specific aims. All of the tools described in the following paragraphs were run using scripts. The programming language Java was also used to implement a variant caller. Details are provided in chapter 3.

The Illumina sequencing platform is distributed with its own sequencing software pipeline (Illumina, 2015d). For the purpose of our analyses, the use of this pipeline is limited to the steps immediately following sequencing: image analysis, base calling and file conversion, the latter including demultiplexing where applicable. The two first steps do not fall within the remit of my data analysis, but awareness of the version of the pipeline used is vital. The data produced at the IGM and GP is provided through the Illumina v1.6 pipeline, while data from the MUGQIC goes through the Illumina v1.8 pipeline. This distinction is important as base call quality is encoded using a different range of ASCII values in either version. Starting with Illumina v1.3 and up to v1.7, the base quality score was encoded in ASCII with an offset of +64, a format unique to the Illumina platform. With Illumina v1.8, the offset changed to +33 to coincide with the Sanger format (Illumina, 2012a). The similarity between encodings has been the source of much confusion and potentially experimental error (Cock *et al.*, 2010). Alignment tools such as BWA (Li and Durbin, 2009) provide an option to convert the illumine-specific encoding format into the more broadly accepted Sanger format (Cock *et al.*, 2010). Failure to explicitly signal the use of the illumina-specific encoding will result in the incorrect interpretation of base quality scores, with negative repercussions on alignment and variant calling.

Sequence data from a single run of the Illumina GAIIx (Illumina, 2012b) is delivered from Illumina’s base calling software in the QSEQ file format (Illumina, 2012a). The QSEQ files are divided by lanes and read direction; given paired-end sequencing. For sequencing runs performed at the IGM and GP, I converted the QSEQ files into the widely-used FASTQ format (Cock *et al.*, 2010) using GERALD, a tool included in Illumina’s data analysis pipeline (Illumina, 2015d). For each sequence lane, two FASTQ files were produced, each containing reads sequenced from the same end. Where sequence data for a single individual was present in multiple lanes, I concatenated the corresponding files so that the remaining pairs of FASTQ files all corresponded to a single sample. The demultiplexed sequence data delivered from the MUGQIC was sent in the BAM file format (Li *et al.*, 2009a). This file provides the required data in compressed form, with sequence alignment. To ensure consistency between samples in my analysis, this alignment was disregarded. I extracted unaligned sequence data from the BAM files (Li *et al.*, 2009a), producing two FASTQ files; given paired-end sequencing, for each BAM file. This was achieved using the program Picard (The Broad Institute, 2015c).

A number of data analysis tools were used repeatedly across studies. For sequence alignment I used BWA v0.7.4 (Li and Durbin, 2009). BWA takes a FASTQ file as input and produces an alignment in BAM format (Li *et al.*, 2009a). At this stage, encoding that was not in the Sanger format was flagged up using the corresponding option in BWA (Li and Durbin, 2009). After every sequence alignment, duplicates were removed using Picard (The Broad Institute, 2015c).

Three different variant callers; two widely-used callers and one of my own devising, were used:

- SAMtools v0.1.18 (Li *et al.*, 2009a)
- GATK UnifiedGenotyper v.2.2.9/ v.3.1.1 (DePristo *et al.*, 2011)
- BAMily (*not yet published*)

The above variant callers can take one or more aligned BAM files, producing a list of variants in VCF file format (Danecek *et al.*, 2011). For each variant caller, a different approach was used. Using SAMtools, I called variants from parent-offspring trio samples together, while producing individual calls for unrelated samples. Before using GATK's UnifiedGenotyper, I used GATK's realignment and recalibration tools on each sample producing recalibrated BAM files. I then called the entire set of recalibrated samples with GATK's UnifiedGenotyper, producing a single VCF output for all the samples of a single study. The use of BAMily is described in chapter 3 and in chapter 4.

For the purpose of both sequence alignment and variant calling, I used the UCSC genome hg19 assembly (UCSC, 2015) in FASTA format. This assembly was used for all analyses described in this thesis. Unless stated otherwise, all variant call data were filtered for the target regions from the SureSelect^{XT} Human All Exon 50Mb kit V4 from (Agilent Technologies, 2015) Agilent Technologies, USA.

Every variant call is accompanied by a quality score (Q) in Phred format; where the error probability is expressed as a score through a logarithmic transformation (Ewing and Green, 1998; Ewing *et al.*, 1998). In chapter 4 and chapter 5, I use this score to capture high-confidence calls. For GATK, the minimum confidence threshold for high-quality variants is a score of Q30 corresponding to a 99.9% probability that the call is correct (The Broad Institute, 2015b). I apply this threshold to SAMtools as well. The SAMtools and GATK genotype predictions were used as an additional filtering criterion in chapter 3 and chapter 4. The quality of selected variant calls was also checked using IGV, an alignment viewer (Robinson *et al.*, 2011). Using this program I was able to flag potential false positives variants by inspecting their sequence context.

The studies undertaken for this PhD were focused on identifying putative rare variants which can cause disease. To filter for variants more likely to fit this description, population frequency and functional annotations are required. For this purpose, I used ANNOVAR (Wang *et al.*, 2010). ANNOVAR provides gene, region and population data annotations for each called locus, by simply adding columns to those already existing in the VCF output. This additional information can be used to filter and rank variant call data. Using my own scripts, I filter the data using the available annotation. For every sequencing analysis, I select non-synonymous variants, falling either in an exonic or splice-site region. Known polymorphisms are removed from the list of variants by comparing variant call data to population databases such as the 1000genomes project (Abecasis *et al.*, 2012) and the NHLBI Exome Sequencing Project (ESP) (NHLBI, 2015). Alleles with a minor allele frequency (MAF) ≥1% in either of these databases are filtered out. To filter out additional common variants, some of which are artefacts, I used an in-house list of variants compiling the results of 418 previous WES analyses. A frequency threshold of 1% is also used in this case. In each study, the potential pathogenicity of variants is partly determined by their predicted functional effect. Various programs provide such predictions via ANNOVAR. These include MutationTaster (Schwarz *et al.*, 2010), Polyphen-2 (Adzhubei *et al.*, 2010) and LRT (Chun and Fay, 2009). I use these to either filter or rank potential disease-causing variants.

Beyond this step, each study tends to differ. Each study has its own methodology which I will describe in the methods section of each subsequent chapters.

Chapter 3. Design, development and testing of a new approach to variant calling

3.1. Summary

In this chapter, I present the design and implementation of a new approach to variant calling. The resulting variant caller, BAMily, is designed to identify putative variants in WES data, shared by a subset of individuals, while absent in another subset. This approach to variant calling is particularly applicable to pedigrees in which multiple related individuals are suspected of sharing a variant causing a Mendelian disease trait. It can be described as a two-step detection process. Variants are first detected in a sample pool, using all available reads. Variant detection in the pool is then used to assist the detection of variant presence or absence in each individual. This approach is particularly useful when a variant, well represented in a pool of individuals, is also present in the reads from a particular individual but, due for example to low quality or coverage, is difficult to distinguish from errors. This approach also involves assigning probabilities to undetected variants in single samples of being truly absent or merely unobserved in an individual. The desired outcome is a variant caller which can be used to uncover variants that segregate with a trait exhibited by groups of individuals, out of a larger set of individuals. For example, the variant caller can be part of a NGS study on patients exhibiting a disease trait suspected of having a shared origin in the form of one or more variants. Families presenting disease traits that follow an autosomal dominant pattern of inheritance is one such case. I compare the implementation of this approach against other variant callers using data from individuals in pedigrees that have also been analysed using genotyping microarray. The results show that my approach provides an advantageous balance between sensitivity, specificity and running time that is complementary to existing methods. As a consequence, I use this new approach to study 8 families. Previous analyses failed to establish plausible disease-causative candidates in most of these families. The development of BAMily presents a new opportunity to attempt to resolve these families.

3.2. Introduction

While NGS platforms produce overall high-quality data from which DNA sequences can be determined, library preparation and sequencing involve a range of steps and optical technology that introduce significant variability in the quality of the data (Kircher *et al.*, 2011). Variation in the base call error rate is in large part mitigated by the presence of overlapping reads. However, variants can become indistinguishable from errors for loci supported by fewer reads. Regions with low read depth pose an additional problem as they may reveal only one of two alleles present at a heterozygous locus (Nielsen *et al.*, 2011). A survey of WES projects centred on Mendelian disease has led Gilissen *et al.* (2012) to report that specific mutations underlying disease traits will be identified in only 60% of families. This statement should be treated with great caution as the outcome of a WES project will depend on many factors such as the number of carriers sequenced and the mode of inheritance observed in a family. Nevertheless, what this estimate conveys is the need for improvements in the analysis of WES data. The challenge for variant callers in the past few years has been to extract variants from reads that are difficult to interpret. I will quickly review some of the approaches that have been used to interpret low-quality data below.

3.2.1. Improving variant calling

Addressing the limitations of NGS has motivated the development of a plethora of increasingly sophisticated tools for sequence alignment, variant calling and filtering (Pabinger *et al.*, 2014). Popular variant callers such as SAMtools or GATK take model based approaches and have led to increased rates of variant discovery; particularly when compared to the original read counting method, while keeping the rate of false positives low (Li *et al.*, 2009a; DePristo *et al.*, 2011). Variant callers rely on estimated calling error probabilities. These estimates are generated by the base calling procedures which are platform specific and themselves subject to biases. There have been efforts to correct known biases, such as the increased propensity for base miscall in later sequencing cycles (Altmann *et al.*, 2012). One approach is to train statistical models on sequencing data. This is the solution proposed by the integrated caller in the Atlas2 suite which uses a logistic regression model trained on exome data to call single nucleotide variants (SNVs) and indels (Challis *et al.*, 2012). Challis et *al*. (2012) report that their caller is thus able to mitigate the effects of exome capture such as reference bias and strong variability in depth coverage. Another approach involves the use of known single nucleotide polymorphisms (SNP) in order to re-estimate the quality of sequence data (Altmann *et al.*, 2012). By identifying sequencing mismatches from the reference not previously identified as nucleotide polymorphisms and treating these as errors, it is possible to derive more accurate error rates from the sequence data. This is the concept behind base quality score recalibration, first introduced with SOAP and then implemented by GATK (Li *et al.*, 2009b; DePristo *et al.*, 2011).

Current variant callers have the ability to detect variants across multiple samples. This provides another potential source of improvement for variant calling. At a single locus, the quality and quantity of reads will differ between samples. For a single sample, variant detection can be hindered by the lack of high-confidence reads, the scarcity of read data, or both (Altmann *et al.*, 2012). Samples that present a high yield of high-confidence reads can help with the interpretation of low-confidence reads in another sample. A higher rate of variant detection can also be achieved by factoring in the relationships that bind the individuals from which these samples originate. In the following section I review WES approaches which use the relationship between samples to increase variant detection.

3.2.2. Variant calling involving multiple samples

3.2.2.1. Family-based sequencing

Families provide a context in which rare variants are concentrated in a relatively small set of samples. Family-based studies can be used to uncover both inherited and *de novo* variants. Disease-causing variants that result from *de novo* mutation are identified by sampling family trios or quartets in which a child is alone in exhibiting a disease trait. Sequencing data from parents and siblings provide a set of variants which can be excluded from the child, since they are inherited (Bamshad *et al.*, 2011). This approach to variant discovery has revealed the role that *de novo* variants play in the manifestation of neurodevelopmental disorders (Veltman and Brunner, 2012a). Familial relationships can also be used to identify mutations causing diseases which follow a Mendelian pattern of inheritance. Candidate mutations are identified by sequencing related individuals; healthy and affected, and selecting those mutations that segregate with the disease. Family-based sequencing studies often rely on variant callers such as SAMtools (Li *et al.*, 2009a) or GATK (DePristo *et al.*, 2011). However, recent attempts to improve variant calling in this domain have led to the development of callers specifically tailored to family sequencing.

Family-based studies benefit from pedigree structure which provides a means to relate each sequenced individual to all others. For a single locus, the sequence reads from one individual are used to determine the presence of absence of a variant in the reads of another following Mendelian inheritance. Thus a variant that would typically be difficult to distinguish from error at a single locus in an individual, benefits from a confident detection of the same variant in a parent individual. The relationship between individuals can be used to derive information that either supports or penalizes the genotype predictions that originally arise from an individual's sequence data. This is the principle behind family-based variant callers such as FamSeq (Peng *et al.*, 2013) and PolyMutt (Li *et al.*, 2012); these callers were published while the present work was in progress. Family-based callers exploit shared variation in closely related individuals in order to strengthen variant detection for each single individual. While the idea of using shared variation in this way has primarily been applied to families, it can also be applied more generally to cases where multiple individuals share disease-causing variants without necessarily all being closely related as will be discussed later in the chapter.

3.2.2.2. Pooled sequencing

The costs involved with WGS, though falling very rapidly, remain an important obstacle in the carrying out of large case-control studies (Derkach *et al.*, 2014). While WES sequencing provides a cheaper alternative for studies which involve only a few dozen samples, it remains often too costly for studies involving several hundreds of samples. Even during the period of this PhD, costs for WGS and WES data generation have substantially converged; the downstream data processing and storage requirements, however, remain significantly different between the two approaches. The demand for cheaper large-scale sequencing studies has motivated the development of pooled sequencing (Derkach *et al.*, 2014). Pooled sequencing consists in creating a DNA library for sequencing from the pooled DNA of all the individuals of a study (Chen and Sun, 2013). Two population sequences; for cases and controls, are effectively created from which the various frequencies of different variants can be estimated.

This approach focuses on establishing the presence of variants in a given pool without assigning these to particular individuals. Variants or clusters of variants likely to play a role in the disease under study will be overrepresented in the case pool compared to controls. This approach compensates for the variability in sequence data quality by using all reads covering a single locus. Variant callers tailored to pooled sequencing include SNVer (Wei *et al.*, 2011), vipR (Altmann *et al.*, 2011) and CRISP (Bansal, 2010). One of the pitfalls of such an approach however, is that individuals can no longer be distinguished. This type of sequencing would therefore not be sufficient to arrive at an individualised variant detection. However, such an approach could be used to uncover variants as a preliminary step to per individual variant calling.

In recent years however, with cross-collaborations providing the means to carry out large studies, pool sequencing has been phased out in favour of sequencing individuals in large cohorts. The study presented in chapter 5 is one example.

3.2.2.3. Somatic sequencing

Population and family-based studies compare sequencing data emanating from multiple individuals. In cancer sequencing, the method is adapted to compare data originating from a single individual, but different cells (Meyerson *et al.*, 2010). DNA samples obtained from tumour cells are compared with samples from non-tumorous somatic cells. As with pooled sequencing, sequencing data are produced for multiple DNA samples. Due to inevitable admixture of different cell populations, both cancer and cancer-free, mutations may manifest as a variant of low frequency. A somatic variant caller, tailored to address the particularities of somatic mutations, is thus required (Pabinger *et al.*, 2014). A few example of these are Mutec (Cibulskis *et al.*, 2013), Strelka (Saunders *et al.*, 2012), SomaticSniper (Larson *et al.*, 2012) and Varscan2 (Koboldt *et al.*, 2012).

3.2.3. Aim

In this chapter, I present a novel approach to variant calling that derives its inspiration from some of the approaches previously described. The variant caller is focused on identifying variants that are predicted to be shared by a subset of individuals while absent in another subset. In the context of a disease trait, these subsets correspond to the pattern of inheritance exhibited by the trait. By changing the scope of interest from all variants present across the set of individuals to a more meaningful subset of shared variants, we can tackle detection of disease-causing variants from a different angle. The caller first identifies variants in the pool of individuals we wish to analyse. The pool will be strongly enriched for shared variants. In this step, no distinction is made between individuals. Variant detections in the pool provide data to make inferences about the existence or absence of variants in distinct individuals, leading to variant re-discovery in reads previously too difficult to call. Using this approach we can also estimate a probability of non-discovery of a true variant or consolidate its absence at loci where no detection has occurred. Many variants are likely to be excluded from studies due to poor sequence coverage in one or two individuals, resulting in failure to identify causative alleles in WES and WGS studies. My approach addresses this problem by estimating the probability that a variant is present even if not observed in an individual's reads and incorporating this probability in the final variant calls assigned.

3.2.4. Presentation of the 8 families analysed

Using the new approach to variant calling presented in this chapter, I analysed 8 families, in which several members share a disease phenotype. These families were part of a larger effort to identify disease-causative variants in families in which the disease trait followed a Mendelian pattern of inheritance. The 8 families correspond chiefly to cases unresolved at the time of study that had already undergone multiple analyses of WES data. With its new approach, BAMily provided the opportunity to analyse these again. The principal investigators with clinical responsibility for each family proposed possible modes of inheritance based on family segregation. In several of these families, the pattern of inheritance of the disease strongly suggested either an autosomal dominant or autosomal recessive mode of inheritance with full penetrance. In other families, some patients show what could be a dominant disease, but with reduced penetrance. Based on this observation and patient availability, a number of individuals in each family were sequenced. In Figure 9 to 16, I provide the corresponding 8 pedigrees and the justifications that we used to assign specific patterns of inheritance conjectured.

Family 1:

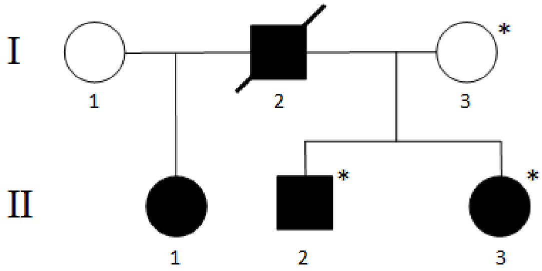

Figure 9. Four individuals exhibit hereditary scleroscing poikiloderma (HSP) with tendon contracture, myopathy and pulmonary fibrosis (Mercier *et al.*, 2013). The disease is assumed autosomal dominant. *Individuals sequenced.

Several members of the first family exhibit a very rare disease, hereditary scleroscing poikiloderma (HSP) accompanied by muscle defects; tendon contractures and myopathy, as well as scarring of the lungs; referred to as pulmonary fibrosis (Mercier *et al.*, 2013). The corresponding pedigree is shown in Figure 9. Mercier *et al*. (2013) describe HSP as "a combination of mottled pigmentation, telangiectasia, and epidermal atrophy in the first few months of life". In this context, telangiectasia refers to the visible dilation of capillaries on the surface of the skin. Describing the pedigree presented above, Khumalo *et al*. (2006) proposed that the disease follows an autosomal dominant pattern of inheritance. Two siblings and the unaffected mother were sequenced.

Individuals in the extended pedigree, reproduced in Figure 10, suffer from atrioventricular septal defect (AVSD) and, in two cases, ostium primum atrial septal defect (ASD). ASD and AVSD both consist of a gap in the wall that separates left and right structures of the heart. Both the atria and ventricles are involved in AVSD, while the atria only are involved in ASD (Kaza *et al.*, 2013). Despite the phenotypical differences, the development of isolated AVSD shares an embryological mechanism with ostium primum ASD, both originating from the abnormal development of endocardial cushions that separate the different chambers of the normal heart (Kaza *et al.*, 2013). Therefore, family members with either AVSD or primum ASD were considered to have equivalent phenotypes in this pedigree. According to D'Allesandro *et al.* (2015), approximately 30% of AVSD are attributable to either chromosomal or single-gene defects. In this family, the segregation of CHD, suggests that an autosomal dominant variant with partial penetrance is causing the reported septal defects. A total of five individuals; two obligate carriers, two individuals with AVSD and one with ASD, were sequenced.

Family 2:

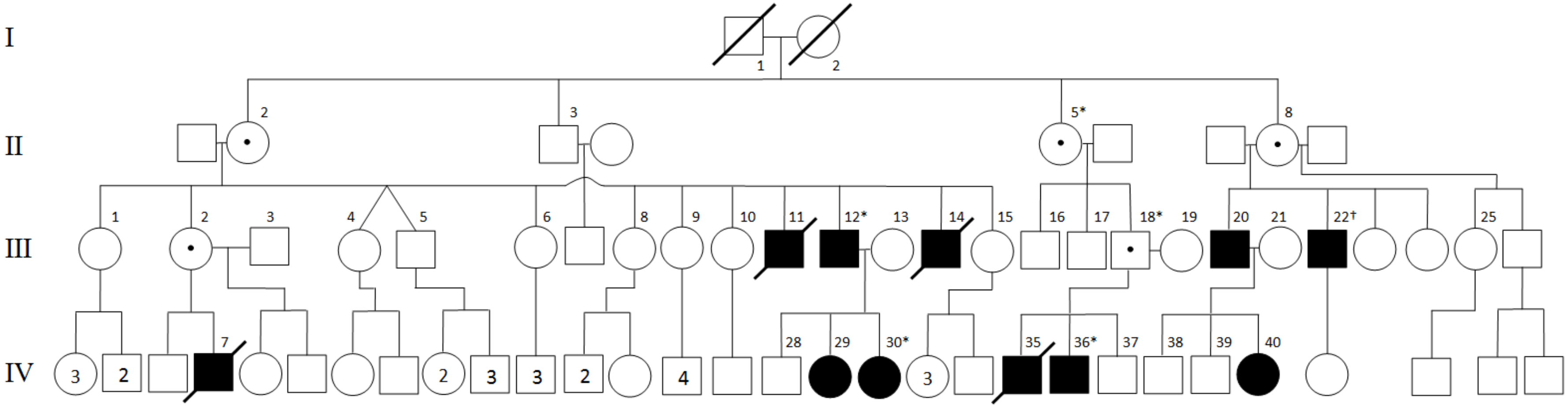

Figure 10. Multiple individuals across an extensive family tree present atrioventricular septal defects (AVSD). Disease assumed to be caused by a single variant with reduced penetrance. *Individuals sequenced. †A DNA sample that does not match this individual was sequenced due a due to sample mix-up. The sequence data was therefore not retained.

Family 3:

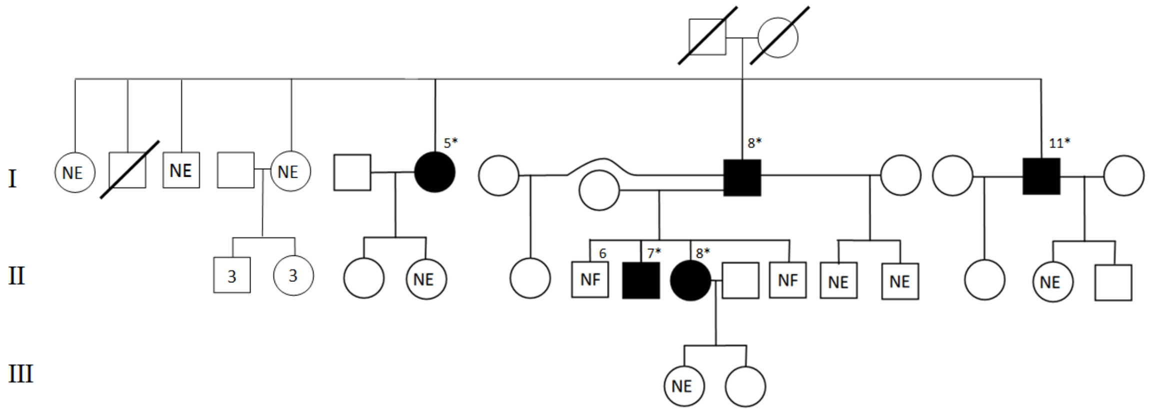

Figure 11. The disease trait observed in five individuals has been described as atypical Brugada syndrome. Disease is assumed autosomal dominant. Individuals with a normal electrocardiogram are marked NE. Individuals that did not present the syndrome as a result of Flecainide stimulation are marked NF. *Individuals sequenced.

Family 3 includes several cases of atypical Brugada syndrome (BrS). The phenotype observed in the index case of this family was ventricular fibrillation, where cardiac muscles of the ventricles contract asynchronously, manifesting as cardiac arrest which without prompt defibrillation would prove fatal. The corresponding pedigree is shown in Figure 11. BrS describes a change in normal electrical activity of the heart during a particular segment of a heartbeat (Antzelevitch *et al.*, 2005). The electrocardiography of patients with BrS reveals an elevation of the ST-segment, abnormal heart rhythm; arrhythmia, and sudden cardiac death (Antzelevitch *et al.*, 2005). Not all carriers of the syndrome show this pattern at resting heart rate. Flecainide can be used to unmask this pattern in disease carriers. Whether the patterns were seen at baseline or after injection of Flecainide, the clinical examiners determined that the current cases show an electrophysiology that deviates from the usual presentation for Brugada and therefore identified it as atypical. BrS is typically described as an autosomal dominantly inherited disease, with a total of 16 genes currently associated with the syndrome, the most common being *SCN5A* in which loss-of-function mutations are found in about 20% of cases (Brugada *et al.*, 1993). However, studies in families harbouring *SCN5A* mutations have shown low disease penetrance, and recent data have served to further emphasise the genetic complexity of the phenotype; for example Bezzina *et al*. (2013) showed evidence for association of common alleles at *SCN5A* and *HEY2* with BrS.

Family 4:

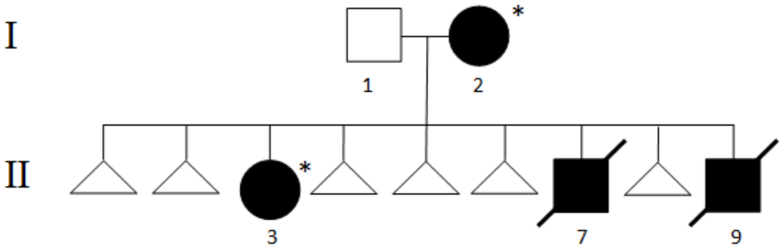

Figure 12. Mother and offspring show discordant CHD phenotypes along with other defects. Six pregnancies did not come to term. The disease pattern of inheritance is unclear. *Individuals sequenced.

In the family, represented in Figure 12, a mother and her offspring present complex cardiac defects along with a number of other pathologies. The mother has a patent ductus arteriosus, a VSD, aortic regurgitation, and a common brachiocephalic trunk. She is missing the gall bladder. At 14, she also developed lymphoedema, a defect of the lymphatic system that causes tissue swelling with accumulation of lymph fluid (Cemal *et al.*, 2011). Additionally, the mother has coeliac disease, an autoimmune disease that leads to gluten intolerance (Leeds *et al.*, 2008). The surviving child has dextrocardia; the major axis of the heart is orientated towards the right mirroring the position of a normal heart (Bernasconi *et al.*, 2005). Other heart defects include VSD and an interrupted aortic arch (IAA). Another child, who died during intrauterine gestation, displayed a common arterial trunk, ASD, subvalvular VSD and ventricular hypertrophy. The presence of bilaterally tri-lobed lungs suggests laterality disturbance is responsible for the observed cardiac malformations. Another child, who died soon after birth, exhibited VSD, abnormal lung lobation, a small adrenal gland and spleen and hypospadias; where the urethra is misplaced. This child is also characterised by the absence of a gall bladder. The mother-daughter pair was sequenced.

Family 5:

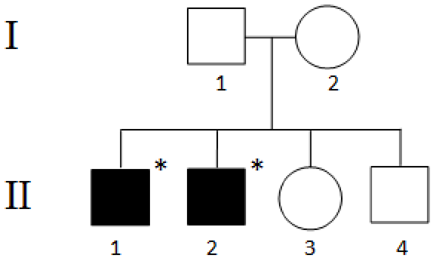

Figure 13. A sibling pair, both sequenced, exhibit double-outlet right ventricle. A number of possible modes of inheritance are possible for this family. *Individuals sequenced.

In the family depicted in Figure 13, a pair of siblings was found to share double-outlet right ventricle (DORV). DORV describes a range of defects that involve the aorta and pulmonary artery both arising from the right ventricle (Obler *et al.*, 2008). The results of a study by Obler *et al*. (2008) suggest that a little under half of cases of DORV arise as a result of chromosomal abnormalities. Beyond chromosomal abnormalities, a number of SNVs and indels have been implicated in the development of DORV, both in humans; for example in *CFC1*, and knockout mice (Obler *et al.*, 2008). In this particular family, the parents and two other siblings do not exhibit the disease. A number of modes of inheritance can be hypothesised for this family. Only the sibling pair with DORV has been sequenced. This was justified by the absence of a clear disease-carrying status for parents.

Family 6:

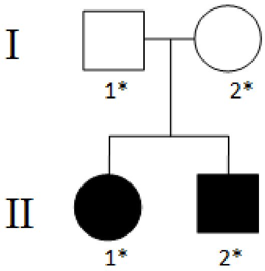

Figure 14. A sibling pair in a nuclear family exhibit major aortopulmonary collateral artery with pulmonary atresia. One patient (II:1) presents a number of other cardiac defects. Neither parent presents a disease phenotype. A number of modes of inheritance are possible for this family. *Individuals sequenced.

In the family represented in Figure 14, the two siblings exhibiting major aortopulmonary collateral artery (MAPCA) with pulmonary atresia (PA). Additionally, the female sibling has a VSD, a right-sided aortic arch and a left-sided innominate vein. Pulmonary atresia with VSD and MAPCA is typically considered a variant of the tetralogy of Fallot phenotype (Prieto, 2005). MAPCA describes the persistence of arteries branching off from the aorta and supplying the pulmonary system in the event of an underdevelopment of the pulmonary valve and arterial system (Boshoff and Gewillig, 2006). MAPCA is often the result of PA which occurs when the pulmonary valve leaflets are fused shut (Bailliard and Anderson, 2009). While neither parent exhibited the disease, the presence of MAPCA in both offspring suggests that some genetic factor could be causing the disease. To explore this possibility, all four individuals were sequenced.

Family 7:

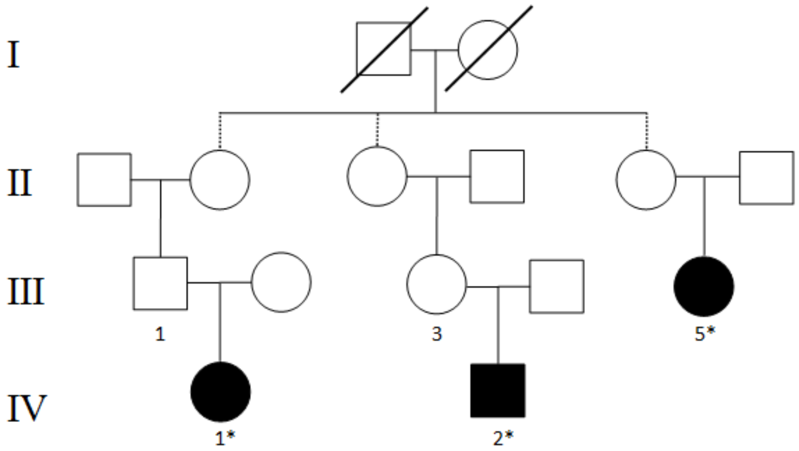

Figure 15. Three related cousins, each exhibit Tetralogy of Fallot. The manifestation of ToF in cousins suggests some genetic variant with partial penetrance could underlie these cases. *Individuals sequenced.

Three distantly related cousins; represented in Figure 15, were found to each exhibit Tetralogy of Fallot (ToF). A description of ToF is given in Chapter 1 and further details provided in chapter 5. Each of the three patients was initially treated separately and then discovered to be related. Using linkage analysis to determine segments of DNA identical by descent, it was determined that III:1, III:3 and III:5 were cousins (personal communication). The recurrence of ToF in three closely related individuals suggests the possibility of some genetic variant with reduced penetrance underlying the disease.

Family 8:

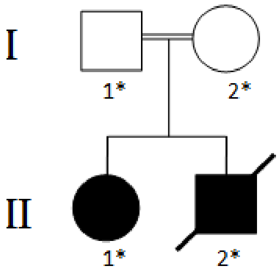

Figure 16. A sibling pair; one of which is deceased, have relapsing cardiomyopathy. Parents have been found to be consanguineous. As a consequence, the disease is assumed autosomal recessive. *Individuals sequenced.

In Family 8, a sibling pair exhibits relapsing dilated cardiomyopathy. Neither parent is affected. Crucially however, the parents are consanguineous; they are first cousins. This suggests that the disease is autosomal recessive. The pedigrees are reproduced in Figure 16. The cardiomyopathy in this family is described as relapsing as both affected members have experienced heart failure on multiple occasions. Cardiomyopathies are diseases of the heart muscle, encompassing a number of structural and functional defects (Sisakian, 2014). Dilated cardiomyopathy, the most common form of cardiomyopathy, is characterised by either the left or both ventricles being dilated and, consequently, dysfunctional (Sisakian, 2014). Up to 40% of dilated cardiomyopathies are found in Mendelian cases (Sisakian, 2014). All four members of the nuclear family were sequenced. Also of note, both affected siblings suffered intractable epilepsy and had significant learning difficulties. The male sibling died of heart failure. The female sibling after two relapsing episodes requiring support with a cardiac Ventricular Assist Device (VAD) and Extracorporeal Membrane Oxygenation (ECMO) received a heart transplant.

The study of these 8 families is the first WES study to incorporate this new approach to variant calling. This chapter therefore not only covers the design, implementation and testing of this approach; through BAMily, but also presents its first application.

3.3. Methods

I devised a novel approach to variant calling. To implement this new approach, I used the Java programming language. The program takes in a set of samples in the BAM file format (Li *et al.*, 2009a) with a carrier status applied to each. The status allows the variant caller to distinguish samples in which we expect to see a mutated allele from those in which no such allele should be found. The program also allows users to set a number of variables, such as the number of independent alleles expected in the pool based on the relationship between individuals. Decryption of the binary-encoded BAM files is handled by the SAM-JDK library (The Broad Institute, 2015c). The calls that are generated by a variant caller are stored in files that follow the VCF file specification (Danecek *et al.,* 2011) although a few columns are specific to this implementation. The approach was implemented in a program: BAMily. The accuracy of calls made by this implementation was then tested.

3.3.1. Measures of shared variant detection quality

The value of our approach and its implementation can be measured in its ability to correctly identify variants shared by a predetermined subset of individuals that are absent in another subset. Both sequencing artefacts and variants that do not segregate with the predetermined pattern need to be excluded and are therefore considered negatives. A comparison of known genotypes with the variants presented by our caller allows us to estimate sensitivity and specificity of the program. In this context, sensitivity corresponds to the rate with which the caller identifies variants that are known to follow the desired pattern of inheritance. Let $F = T + M$ be to the number of variants known to follow the desired pattern of inheritance, where T corresponds to variants identified by the variant caller and M corresponds to those missed by the caller. We have:

$$\boldsymbol{Sensitivity = \frac{T}{F}}$$

Specificity corresponds to the rate with which the caller excludes variants that do not follow the desired pattern of inheritance as well as genotype mis-assignments resulting from sequencing artefacts. Let $N = E + P$ correspond to the number of variants that are known not following the desired pattern of inheritance, where E corresponds to variants correctly excluded by the variant caller, while P corresponds to those incorrectly detected by the caller. We have:

$$\boldsymbol{Specificity = \frac{E}{N}}$$

Data produced using genotyping microarray provides a list of SNPs present in a single individual. Data from the individuals used for sequencing can be used to estimate the sensitivity and specificity of different variant callers. While it is true that microarray data are not devoid of error, it is sufficiently accurate for the purpose of estimating sensitivity and specificity, as was first published by Ng et al. (2009). Positions were separated between those that carried alleles that that followed the pattern of interest and those that did not. The former was used to estimate sensitivity while the latter was used to estimate specificity. Any positions not represented in the microarray were excluded.

Exome sequencing and microarray data were obtained for 21 complete parent-offspring trios and a set of five additional individuals; three first cousins, their mutual uncle and one unrelated individual, from previous experiments. The 21 trios were obtained from a study of Tetralogy of Fallot (ToF) in which the offspring in each trio is the affected proband. Sequencing was performed on the Illumina GAIIx (Illumina, 2012b) with target enrichment for exome capture using SureSelect^{XT} Human All Exon 50Mb kit (Agilent Technologies, 2015). The Infinium HD Human660W-Quad Beadchip (Illumina, 2015a) was used as genotyping microarray. The chip has markers for 657,366 SNPs and copy number variants (CNV) spanning the entire length of the genome. For each of the 21 trios, around 13,726 positions represented in the microarray were within sequencing targets.

The microarray data did not always provide a clear genotype assignment. For each trio, positions missing a genotype assignment were excluded from the evaluation. There are therefore some small differences in the number of positions represented per trio. In order to measure the effect of sample set size on caller performance, I used five samples; four originating from a single family, which were included in a large quantitative genetic study of hypertension. Figure 17 shows the corresponding pedigree for the 4 related individuals. In this case, the sequencing was performed on the Illumina HiSeq 2500 (Illumina, 2014b). Samples were aligned using BWA v0.7.4 (Li and Durbin, 2009). For the five additional individuals, genotypes were available for 14,087 positions. Differences between these two sets are largely due to differences in the sequencing platform used.

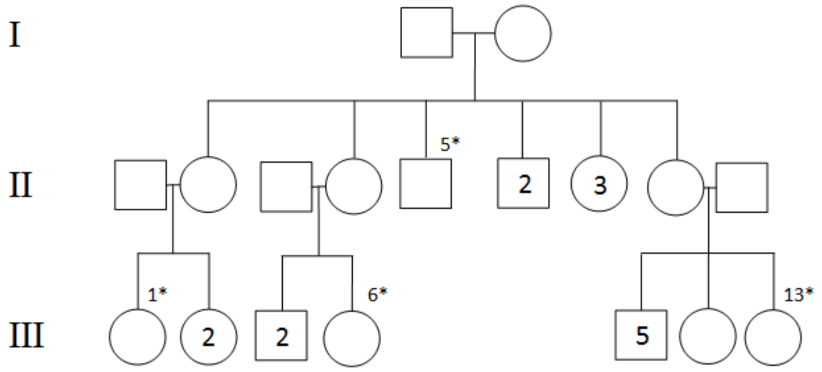

Figure 17. Pedigree of the four related individuals included in my analysis. Samples were originally sequenced as part of a hypertension study. *Individuals sequenced.

First I assessed variants shared between a father-child pair and absent in the mother in each of our trios. This corresponds to the variant calling step that would be used in the search for a paternally-inherited disease-causing variant. To do this, I derived a sensitivity and specificity measurement for each trio. This measurement was performed using five different variant callers: SAMtools v.0.1.18 (Li *et al.*, 2009a), GATK UnifiedGenotyper v3.1.1 (DePristo *et al.*, 2011), GATK+FamSeq, GATK+PolyMutt and BAMily. FamSeq v.1.1.0 (Peng *et al.*, 2013) and PolyMutt v.0.1.5 (Li *et al.*, 2012) used the genotype likelihoods produced by GATK. For each caller, with the exception of BAMily, I filtered for variants shared by father-child pair and not by the mother using their respective genotype calls. This last step is already integrated into BAMily's design.

I then made a series of sensitivity and specificity measurements on the additional five individuals, investigating the effect of the number of individuals analysed and their relatedness on sensitivity, specificity and the number of variants produced. Starting with three samples, I first singled out variants shared between two related individuals and absent in the unrelated individual, getting an average sensitivity and specificity from each possible arrangement of samples. The same was done with four and five samples, each time searching for variants shared by the individuals that were related. Following this, using sample from all five individuals, I performed every possible combination of assuming non-carrier status of one, then two, of the related individuals and obtained sensitivity and specificity measures. The unrelated individual here serves to exclude variants that are not family-specific and to prevent common sequencing artefacts from being detected as variants.

3.3.2. Application of BAMily to the WES study of 8 families

The WES analysis of the 8 families follows some of the methods outlined in Chapter 2. As in other studies, sequencing was performed at the Institute of Genetic Medicine (IGM) using the Illumina Genome Analyzer IIx (GAIIx) (Illumina, 2012b). Target enrichment for exome capture was performed using SureSelect^{XT} Human All Exon 50Mb kit V4 (Agilent Technologies, 2015) from Agilent Technologies, USA. I used BWA v0.7.4 (Li and Durbin, 2009) for sequence alignments, Picard (The Broad Institute, 2015c) for duplicate removal and BAMily for variant calling. In the analysis of each family using BAMily, I added one individual from Family 1 as non-carrier (I:3, see Figure 9) to better exclude any systematic errors that might be shared between individuals and therefore preclude the possibility of these errors being called as genuine shared variants. For comparison, I also called variants using SAMtools v0.1.18 (Li *et al.*, 2009a). SAMtools assigns a genotype to each individual for every variant site. Variants were selected if the corresponding genotype assignments fit the pattern of inheritance agreed upon on at the outset of the analysis. This step was not required for BAMily, with the pattern of inheritance being set as part of the variant calling. For either caller, I excluded variants not called with high-confidence (Q30). I used the online implementation of Annovar (Chang and Wang, 2012) to access the latest version of annotation tools described in Chapter 2. In this particular case, I incorporated the ExAC (The Broad Institute, 2015a) population database into the analysis. Using a criterion of exclusion lower than that outlined in Chapter 2 (MAF≥1%), when filtering against the 1000genomes risks the exclusion of rare pathogenic variants, the database being based only on 1,092 genomes of variable coverage, and thus the estimates of allele frequencies >1% being subject to considerable error (Abecasis *et al.*, 2012). On the other hand, ExAC aggregates exome sequencing from 60,706 unrelated individuals meaning that alleles with frequencies >0.5% can be confidently recognised from the dataset as polymorphisms (The Broad Institute, 2015a). I therefore excluded any variant with a MAF>0.5% in ExAC. I also filtered the list of variants against an in-house list of variants already found in 418 previous WES analyses. This step excludes variants that have been uncovered before in studies of different conditions undertaken at our institute. Those analyses used the same laboratory methods and similar bioinformatics pipeline, and are therefore likely to represent systematic errors rather than disease-causing variants of interest. MutationTaster (Schwarz *et al.*, 2010), Polyphen-2 (Adzhubei *et al.*, 2010) and LRT (Chun and Fay, 2009) were used to assess potential pathogenicity. The resulting list of variants was evaluated using resources such as the Online Mendelian Inheritance in Man (NCBI, 2015) and GeneCards (Rebhan *et al.*, 1998) compendiums. These results provided information on the gene context for each variant. I also referred back to results from previous analyses, performed either by me or colleagues. I used the integrative genomics viewer (IGV) (Robinson *et al.*, 2011) in order to inspect each variant's sequence context. This tool was invaluable in picking out probable artefacts in low complexity regions and among indels. I selected probable disease-causing variants for validation if they appeared in genes which could conceivably play a role in the disease. For example, for a CHD phenotype, variants appearing in genes expressed in the embryonic heart or genes leading to a cardiac defect in mouse models were retained. I report the variants that were sent for validation in Section 3.4.5. I also used the opportunity presented by this analysis to look for CNVs. For this purpose, I used ExomeDepth (Plagnol *et al.*, 2012). I looked for CNVs found across all putative disease variant carriers in a family that stretched across at least 0.1Mbs. The confidence with which each CNV call is made is given with a Bayes factor (BF). I selected CNVs called with high confidence (BF≥30). ExomeDepth requires controls, ideally with the same overall read depth. Sequence data from 20 of the parents of probands with transposition of the great arteries (TGA); presented in Chapter 4, were used as controls.

3.4. Results

In this section I first present the design for my new approach to variant calling which I implemented as the variant caller BAMily. I then present the results of multiple performance tests which provide a means of assessing BAMily in the context of other widely-used variant caller. Finally, I report the results of the integration of BAMily in the WES studies of 8 families.

3.4.1. A new approach to variant calling

Let us assume that we have a set of sequenced individuals and some putative variant we hypothesise to be present in some individuals while absent in others. Let n be the number of individuals in which that variant is present and m the number of individuals in which it is absent. Let T represent the calling results from the entire pool of samples. For that same locus, let S$_i$ represent the calling results for individual i, such that $S = (S_1, \ldots, S_{n+m})$. Let α be the status for a variant in the pool and β_i the status for a variant in individual i such that $\beta = (\beta_1, \ldots, \beta_{n+m})$. At each locus, the presence of a variant is given as D while the absence of a variant is signalled by $\overline{D}$. For each locus, we want to know the probability a variant is detected in n individuals and not detected in m individuals. My approach can be summed up as the following expression:

$$\begin{aligned}\Pr(\beta_1 = D_1, \ldots, \beta_n = D_n, \beta_{n+1} = \overline{D}_{n+1}, \ldots, \beta_{n+m} = \overline{D}_{n+m} | S, \alpha, T)\\ = \prod_{i=1}^{n} \Pr(\beta_i = D_i | S_i, \alpha, T) \prod_{j=n+1}^{n+m} Pr(\beta_j = \overline{D}_j | S_j, \alpha, T)\end{aligned} \tag{1}$$

I use Bayes' theorem to determine the probability that a variant is present in individual i, given sequence data; from both the individual and pool, and the variant detection status in the pool of samples:

$$\Pr(\beta_i = D_i | S_i, \alpha, T) = \frac{\Pr(S_i | \beta_i = D_i) \sum_{\alpha \in (D, \overline{D})} \Pr(\beta_i = D_i | \alpha) \Pr\,(\alpha | T)}{\sum_{\beta_i \in (D_i, \overline{D}_i)} Pr(S_i | \beta_i) \sum_{\alpha \in (D, \overline{D})} Pr(\beta_i | \alpha) Pr\,(\alpha | T)} \tag{2}$$

Conversely, the probability that a variant is not present in an individual given the data is:

$$\Pr(\beta_i = \overline{D}_i | S_i, \alpha, T) = 1 - Pr(\beta_i = D_i | S_i, \alpha, T) \tag{3}$$

The characteristic feature of my method's two-step nature is the dependence of the variant status β_i on the pool's variant status α.

Let us assume that for a variant to be detected in the pool, it must be present in at least one individual in that pool. Following this assumption, a variant will be present in at least one allele within a pool of f independent alleles. The probability of detecting a variant in a single individual *i* therefore depend on detection in the pool. The probability of a variant in individual *i* given its presence in the pool is as follows:

$$\Pr(\beta_i = D_i|\alpha = D) = \frac{2}{f} \tag{4}$$

And the probability of absence is as follows:

$$\Pr(\beta_i = D_i|\alpha = D) = \frac{f-2}{f} \tag{5}$$

We assume that it is not possible for a variant absent in the pool to be subsequently present in any single individual. We can therefore make the following assertions:

$$\Pr(\beta_i = D_i|\alpha = \bar{D}) = 0 \tag{6}$$

$$\Pr(\beta_i = \bar{D}_i|\alpha = \bar{D}) = 1 \tag{7}$$

Given the calling results T, the probability of presence in the pool can also be obtained by applying Bayes' theorem:

$$\Pr\ (\alpha = D|T) = \frac{\Pr(T|\alpha = D) Pr\ (\alpha = D)}{\sum_{\alpha\in(D,\bar{D})} Pr(T|\alpha) Pr\ (\alpha)} \tag{8}$$

The probability of absence in the pool is similarly given by:

$$\Pr(\alpha = \bar{D}|T) = 1 - Pr\ (\alpha = D|T) \tag{9}$$

In order to define the prior probability of detecting a variant from the sample pool, we must first define the prior probability of a variant being present at a locus for a pair of individuals. Let that event be defined as I_i for an individual i. A generally accepted approximation for this is 10^{-3} (Li *et al.*, 2008). For a small $n + m$, the prior can be extended to a pool of samples and approximated in the following manner:

$$\Pr(\mathrm{\alpha = D}) = \Pr(I_1, \dots, I_{n+m}) \cong \frac{1}{1000}(n + m) \tag{10}$$

This set of expressions has been implemented in the variant caller BAMily.

3.4.2. Base call accuracy across reads from individual

Equation 2 requires a probability that the sequence data from individual i fits a particular detection assignment. The following equations are based on the genotype likelihood equation proposed for Heng Li's statistical framework for SNP calling (Li, 2011). The sequence data S_i can be described as a set of k reads, with l reads corresponding to reference reads (alternatively, the base assignment with the highest count that is not the variant of interest) and $k - l$ reads corresponding to reads showing the variant. We can describe these reads in terms of their error probabilities such that e_j is the error probability associated with read j. The error probabilities are determined from base and mapping scores in the PHRED format called quality scores. This format, along with the formula required to convert quality scores into error probabilities, are detailed in two papers (Ewing and Green, 1998; Ewing *et al.*, 1998) and will therefore not be described here. How a quality score q_j is obtained from mapping and base quality is shown in Equation 16. From the genotype likelihood equation presented in Li (2011) we derive the following probabilities:

$$\Pr(S_i|\beta_i = D_i) = \prod_{j=1}^{l} e_j \prod_{j=l+1}^{k} (1 - e_j) + \frac{1}{2^k} \tag{11}$$

$$\Pr(S_i|\beta_i = \overline{D}_i) = \prod_{j=1}^{l} (1 - e_j) \prod_{j=l+1}^{k} e_j \tag{12}$$

As with other variant callers, error probabilities below a certain threshold are excluded. In this implementation, for a read j to be counted requires $e_j > 0.5$; although this value can be changed by the user.

3.4.3. Base call accuracy across reads from pool

Equation 8 also requires a probability that sequence data fit with a particular detection assignment. Let E be the error probability associated with all reads containing the variant of interest. We have the following probabilities:

$$\Pr(T|\alpha = D) = 1 - E \tag{13}$$

and

$$\Pr(T|\alpha = \bar{D}) = \frac{E}{3} \tag{14}$$

At each base position, there are 3 possible incorrect base call assignments, the sum total probability of which is given by the conditional error probability. Depending on the true base being interpreted by the sequencer, some erroneous base call assignments will be more or less probable. Given that we cannot know the true base being sequenced, we make the simplifying assumption that each incorrect assignment has an equivalent probability of arising in sequencing leading to Equation 14.

The error probability for all reads containing the variant of interest E is obtained from the sum total of quality scores in the PHRED format once converted into error probabilities (Ewing and Green, 1998; Ewing *et al.*, 1998). We can write this quality score Q as the sum of quality scores for r overlapping reads containing the variant of interest with q_j the quality score for read j*:*

$$Q = \sum_{j=1}^{r} q_j \tag{15}$$

Each base in a read is associated with a base call quality score b_j and a mapping call quality score m_j. The overall quality score q_j combines these two qualities as follows:

$$q_j = b_j + m_j - b_j m_j \tag{16}$$

3.4.4. Caller performance

I measured the running time of each caller on sequence data from five individuals. For five individuals, SAMtools requires approximately 8 hours to run. Given that BAMily reports a subset of all variants, its running time depends on the carrier status assigned to each individual. Running time for five individuals fluctuates between 6-8 hours. Both of these programs were run on a single core. By contrast, GATK's realignment and recalibration steps require an average 6 hours per sequenced individual, using two cores in each case. GATK’s UnifiedGenotyper then requires an additional 8.5 hours running on 4 cores. FamSeq and Polymutt require GATK to run and therefore can offer no running time advantage.

I provide an array of sensitivity and specificity estimates for the five variant callers using the method describes in Section 3.3.1. I first performed measurements on trios, focusing on the shared variants between father and child, excluding all variants shared with the mother by either. From the sensitivity and specificity estimates obtained for 21 distinct trios, I determine an average sensitivity and specificity for each variant caller tested. Summaries of these results are provided in Table 4 and Figure 18 with the standard error.

	SAMtools	GATK	FamSeq	PolyMutt	BAMily
Sensitivity %	95.87	97.98	98.02	97.76	98
Specificity %	99.62	99.91	99.87	99.91	99.91

Table 4. Mean sensitivity and specificity computed over 21 trios compared to microarray data. These results are also reproduced in Figure 18.

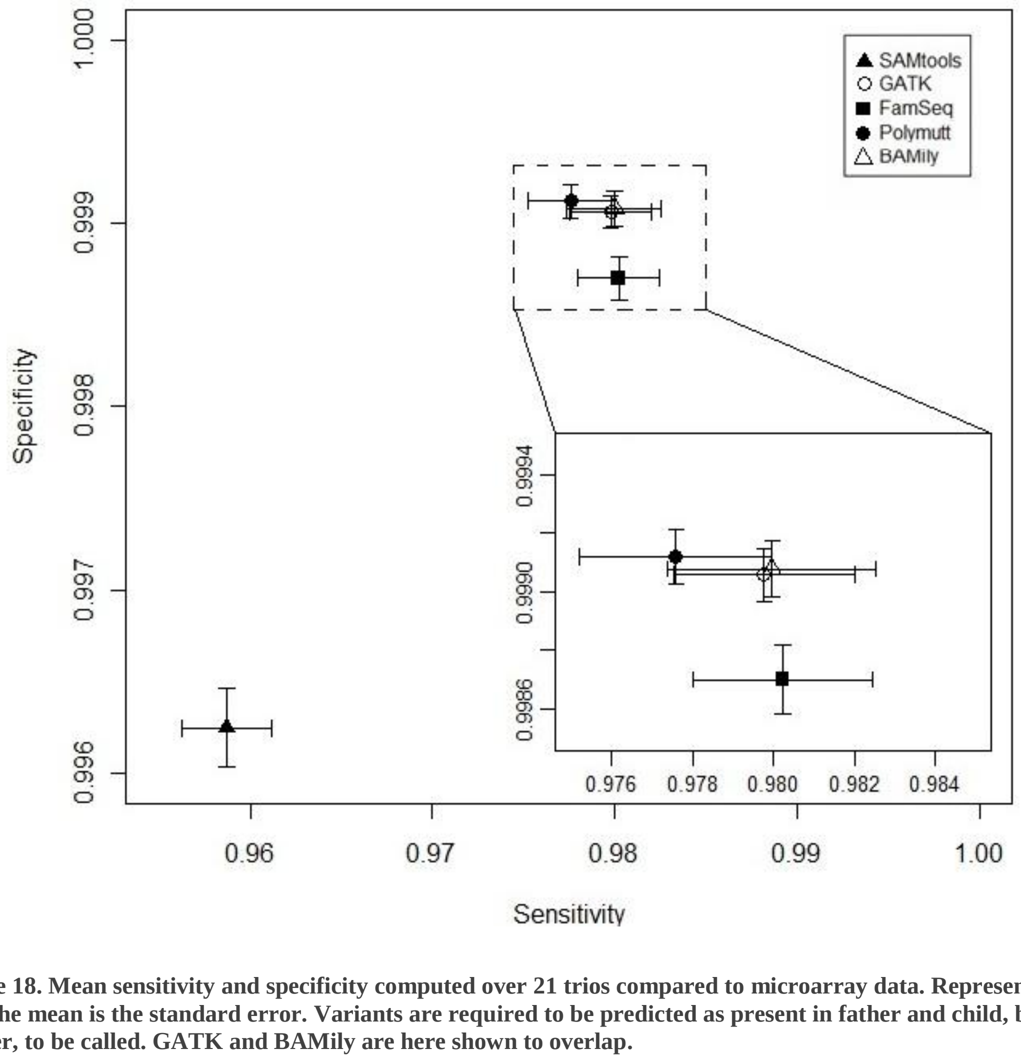

Figure 18. Mean sensitivity and specificity computed over 21 trios compared to microarray data. Represented with the mean is the standard error. Variants are required to be predicted as present in father and child, but not mother, to be called. GATK and BAMily are here shown to overlap.

Over 21 parent-offspring trios, the combination of GATK+FamSeq presents the highest overall sensitivity. However, this comes at a cost in specificity. Likewise, the combination of GATK+Polymutt presents the highest overall specificity, but this comes at the cost of sensitivity. GATK and BAMily sensitivity and specificity largely overlap. BAMily's average sensitivity appears higher, however the difference is not significant (*p*-value=0.197, paired Wilcoxon signed-ranked test). However, it is important to measure BAMily's performance when a larger sample set size is being used, since BAMily is designed to incorporate relationship information from multiple family members, and trio families would not be anticipated to fully utilise the advantages of this method.

I therefore measured sensitivity and specificity for various sample set sizes using sequence data from three cousins, their mutual uncle and one unrelated individual, as shown in Figure 17. I first focused on variants shared between cousin pairs and cousin-uncle pairs, excluding variants shared with the unrelated individual. I then proceeded to increase sample size by incrementally adding one of the remaining cousins or uncle to each pair. In each case, I identify variants shared between the newly added individual and related individuals from the previous call, excluding once more variants present in the unrelated individual.

As sample size increases, we observe an overall increase in specificity and decrease in sensitivity. Additionally, we observe a decrease in the number of variants produced with an increased sample set size. Increasing the number of individuals in which shared variants are sought effectively reduces the number of candidates that will have to be screened. Within a particular set size, the number of variants produced decreases depending on genetic distance between samples. Samples from cousin trios produce more SNVs than uncle-cousins trios for example. GATK, GATK+Polymutt and BAMily show similarly high specificity while GATK+FamSeq displays the highest sensitivity for any case. As before however, the combination of GATK with either Polymutt or FamSeq leads to a decrease in one of these indicators. As with the parents-offspring trios, BAMily and GATK provide a similar balance between sensitivity and specificity. Results are presented in Table 5 and Figure 19.

Still using the set of five individuals, I analysed the data assuming that one of the cousins or uncle was a non-carrier and produced variant calls excluding variants from that individual and the unrelated individual. This was repeated with each family individual, testing every possible combination and obtaining an average. I then assumed two samples were non-carriers and proceeded in the same manner. I found that specificity increases as the total number of variants produced is further reduced. As expected, the total number of variants produced is higher with genetic distance from the non-carriers. In scenarios where the uncle was assumed to be a non-carrier, fewer variants were produced. Sensitivity and specificity for all callers start to align. However, with a smaller number of variants, one missed variant has a large impact on sensitivity as can be seen in the results shown in Table 6.

	SAMtools	GATK	FamSeq	PolyMutt	BAMily
Variants present in cousin-uncle pair, absent in unrelated individual (3 trios)					
Sensitivity %	96.88	97.79	97.98	97.51	97.73
Specificity %	99.9	99.95	99.83	99.94	99.94
False discovery %	1.05	0.53	1.8	0.62	0.62
…in cousin-cousin pair, absent in unrelated individual (3 trios)					
Sensitivity %	96.74	97.72	97.86	97.51	97.72
Specificity %	99.9	99.95	99.81	99.96	99.95
False discovery %	1.12	0.54	2.14	0.45	0.6
…in uncle-cousins trio, absent in unrelated individual (3 quartets)					
Sensitivity %	96.85	97.48	97.82	97.35	97.52
Specificity %	99.91	99.96	99.83	99.96	99.94
False discovery %	1.5	0.6	2.8	0.69	0.98
…in cousins trio, absent in unrelated individual (1 quartet)					
Sensitivity %	96.7	97.11	97.25	96.97	97.11
Specificity %	99.9	99.97	99.87	99.97	99.94
False discovery %	1.82	0.56	2.35	0.56	1.12
…in uncle-cousins quartet, absent in unrelated individual (1 quintet)					
Sensitivity %	96.66	97.01	97.36	96.84	97.01
Specificity %	99.92	99.97	99.88	99.96	99.96
False discovery %	1.96	0.72	2.81	0.9	1.08

Table 5. Sequence data and microarray data comparison for various sample sizes with cousin-cousin and uncle-cousins arrangements shown separately. In this table, positions refer to the positions covered by the microarray data. For a more complete set of metrics, see Appendix, Table S1.

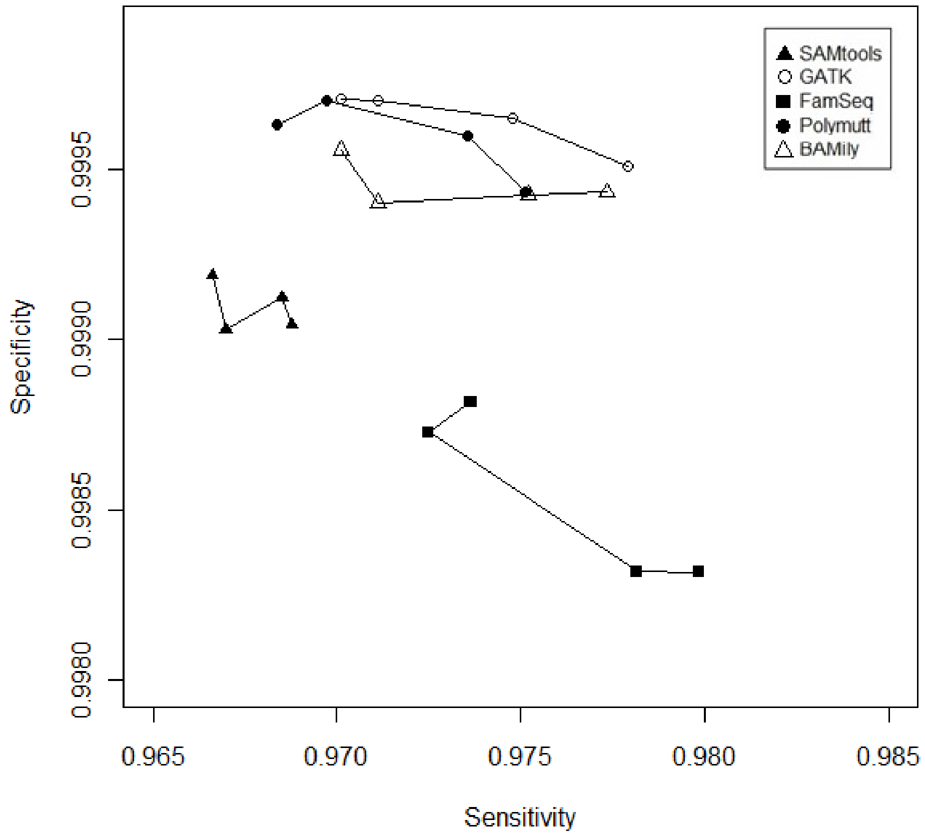

Figure 19. Sequence data and microarray data comparison for various sample sizes (In this case, the average of cousin-uncle pairs is used to illustrate a sample size 2). Each point represents an increase in sample size when reading from right to left; with sensitivity decreasing.

	SAMtools	GATK	FamSeq	PolyMutt	BAMily
Variants in cousins, absent in uncle and unrelated individual (1 quintet)					
Sensitivity %	96.2	96.2	96.2	96.2	96.2
Specificity %	99.97	99.99	99.99	99.99	99.98
False discovery %	2.56	1.3	1.3	1.3	1.94
…in uncle-cousins trio, absent in one cousin and unrelated (3 quintets)					
Sensitivity %	95.99	97.18	97.92	97.18	97.77
Specificity %	99.98	99.99	99.94	99.99	99.98
False discovery %	1.37	0.91	3.64	0.91	1.33
…in cousin pair, absent in uncle-cousin-unrelated trio (3 quintets)					
Sensitivity %	96.52	97.85	97.47	97.85	97.85
Specificity %	99.98	99.99	99.98	99.99	99.99
False discovery %	1.38	0.6	1.38	0.6	1.17
…in uncle-cousin pair, absent in cousins-unrelated trio (3 quintets)					
Sensitivity %	95.38	97.03	97.38	97.18	96.84
Specificity %	99.98	99.99	99.96	99.99	99.99
False discovery %	1.83	0.5	2.65	0.69	0.69

Table 6. Sequence data and microarray data comparison for a fixed sample size of five, assuming different pairs and trios of individuals to be non-carriers. In this table, positions refer to the positions covered by microarray data. For a more complete set of metrics, see Appendix, Table S2.

The measurements shown in Table 4 and Figure 18 as well as Table 5 and Table 6 indicate that BAMily is overall more sensitive and specific than SAMtools, more sensitive than GATK+Polymutt and more specific than GATK+FamSeq. For large sample set sizes ≥4, BAMily closely mirrors GATK's sensitivity as shown in Table 5, occasionally appearing more sensitive. Regardless of the sample size, BAMily's specificity is either matched or lower than that of GATK's. Given the diminishing size of the total variant output with each new sample, a slight decrease in specificity is not particularly disadvantageous.

3.4.5. Analysis of the 8 families

The analysis of 8 families using BAMily and SAMtools revealed a number of variants that appeared to have disease-causing potential. This demonstrates the applicability of the approach implemented as BAMily. Table 7 provides a list of candidates that were sent for validation.

Family ID	**Phenotype**	**Carrier status (genotype for SAMtools)**	**Rare variants**		**Predicted deleterious***		**Candidates retained**
			BAMily	SAMtools	BAMily	SAMtools	
Family 1	HSP	mother non-carrier, offspring (heterozygous)	84	55	23	20	3**
Family 2	ASD/AVSD	All carriers (heterozygous)	25	1	3	1	0
Family 3	BrS	All carriers (heterozygous)	77	18	45	1	1**†
Family 4	Complex CHD	All carriers (not specified)	177	128	50	43	2
Family 5	DORV	All carriers (heterozygous)	175	116	36	39	5**
Family 6	MAPCA	Mother and offspring carriers, father non-carrier/ Father and offspring carriers, mother non-carrier (heterozygous)	69/94	64/74	17/23	27/24	1
Family 7	ToF	All carriers (heterozygous)	114	4	51	2	0
Family 8	Relapsing cardiomyopathy	All carriers (heterozygous in parents, homozygous in offspring)	17	9	7	6	3

Table 7. Number of rare variants detected by BAMily and SAMtools in each family following filtering. *Predicted as possibly delirious (or equivalent) by Polyphen2, MutationTaster and LRT; or an indel. **Candidates retained was not predicted as deleterious by all pathogenicity predictors. † Candidate was not found by BAMily.

Family ID	**Position**	**Nucleotide and amino acid change (transcript)***	**Gene**	**Validated**
Family 1	chr3:19479731	c.G1253A:p.R418Q (ENST00000328405)	***KCNH8***	**Yes**
	chr8: 11412934	c.G713A:p.R238Q (ENST00000259089)	***BLK***	**Yes**
	chr17:11572991	c.T3233G:p.I1078S (ENST00000262442)	*DNAH9*	No
Family 2	No candidates			
Family 3	chr16:88792738	c.C3922G:p.L1308V (ENST00000301015)	*PIEZO1*	No
Family 4	chr4:114276906	c.G7132A:p.E2378K (ENST00000357077)	*ANK2*	No
	chr22:19213864	c.G1825A:p.D609N (ENST00000427926)	*CLTCL1*	In progress
Family 5	chr3:193855858	c.G679A:p.G227S (ENST00000232424)	*HES1*	In progress
	chr5:42700088	c.A602G:p.E201G (ENST00000230882)	*GHR*	No
	chr5:118556760	c.G7679A:p.G2560D (ENST00000311085)	*DMXL1*	No
	chr6:87969514	c.C6167A:p.S2056Y (ENST00000369577)	*ZNF292*	No
	chr6:105609396	c.C389T:p.A130V (ENST00000254765)	*POPDC3*	In progress
Family 6	chr9: 139409980	c.G1858T:p.D620Y (ENST00000277541)	***NOTCH1***	**Yes**
Family 7	No candidates			
Family 8	chr2: 27427777	c.C757T:p.253W (ENST00000310574)	***SLC5A6***	**Yes**
	chr2: 29383269-29383270	c.1470_1471del:p.R490fs (ENST00000320081)	***CLIP4***	**Yes**
	chr5:1681999	c.G1324A:p.D442N (ENST00000332966)	***SLIT3***	**Yes**

Table 8. List of candidates selected for validation work from BAMily and SAMtools analysis. *For brevity, variant is shown for one transcript only.

In Family 1, I rediscovered three missense variants found in a previous analysis in genes *KCNH8*, *BLK* and *DNAH9*. All three were detected by SAMtools and BAMily. LRT was unable to produce a pathogenicity prediction for the variant in *BLK*, but both Polyphen2 and MutationTaster predicted the variant to be deleterious. I sent the candidates for validation and two were found to be genuine variants, as show in Table 8. The variant in *BLK* was of particular interest. A review of the literature on *BLK* reveals that the gene coincides with a region previously associated with keratolytic winter erythema (KWE) (Appel *et al.*, 2002). Beyond the fact that both KWE and HSP are skin disorders, they share similar symptoms such as the high skin pigmentation; or hyperkeratosis, and sclerosis of the hands and feet (Appel *et al.*, 2002; Mercier *et al.*, 2013). A parallel investigation of this family had found evidence that the variant causing HSP in Family 1 was in the gene *FAM111B* (Mercier *et al.*, 2013). The evidence supporting a causative role for the *FAM111B* variant was the presence of two more variants, including one *de novo* variant, in the same gene in four unrelated families exhibiting a similar phenotype

(Mercier *et al.*, 2013). The *FAM111B* variant reported in that paper was not retained in my analysis, as it was predicted to be benign by the pathogenicity predictor MutationTaster. Nonetheless, this variant was detected by both BAMily and SAMtools. Despite these findings, *BLK* could still play a role in the disease, possibly as a phenotypic modifier. Study of the additional families with this very rare disorder would be necessary to test this hypothesis. ExomeDepth did not reveal any CNV candidates.

In Family 2, BAMily and SAMtools found few variants both rare and predicted deleterious. In SAMtools, the only rare variant found to be deleterious by Polyphen2, LRT and MutationTaster had been identified in previous analyses, but had failed validation through Sanger sequencing. BAMily identified 25 rare variants segregating with disease but 15 were predicted to be benign by all pathogenicity predictors. Furthermore, several variants occurred within the same set of genes, suggesting false positives. For example, 13 variants were all found in killer cell immunoglobulin-like receptor (KIR) genes in a region known for being highly polymorphic (Middleton and Gonzelez, 2010). No candidates were retained. No rare CNVs were found to be shared between individuals.

In Family 3, BAMily and SAMtools also found few variants when all filters were applied. BAMily identified 43 rare indels, but most were clearly false positives based on allele frequency in calls and further inspection in IGV. I therefore looked for rare variants predicted to be deleterious by at least one pathogenicity predictor. I selected a missense variant discovered through SAMtools in *PIEZO1* and predicted to be deleterious by MutationTaster and Polyphen2. This variant was not found through BAMily. The gene encodes an ion channel protein which could play a role in vascular development (Ranade *et al.*, 2014). Validation work revealed that the variant was present in only 4 of the 5 sequenced individuals. The clinical genetics team responsible for the care of this family have since discovered a deletion spanning the entire length of *SCN5A* in all carriers (personal communication). *SCN5A* is the principal causative gene for BrS (Brugada *et al.*, 1993). This gene-wide deletion could not have been identified with the SNV-focused methods used in the comparison between variant callers. However, CNV analysis using ExomeDepth did corroborate this finding in all five carriers, with one notable distinction: the CNV was found to span the entire length of two genes rather than one, *SCN5A* and *SCN10A*. The latter has also been implicated in cases of BrS (Hu *et al.*, 2014). Also, although deletions spanning *SCN5A* have been reported in BrS, deletions spanning both

SCN5A and *SCN10A* have not. The cardiac electrophysiology phenotype of the family was noted to be atypical for BrS; whether this reflects haploinsufficiency of both *SCN5A* and *SCN10A* will be the subject of further investigation. The deletion also stretched across the last exon of the gene *EXOG*, a gene previously not thought to be involved in BrS.

In family 4, given the difficulty in ascribing a mode of inheritance for the putative variant, I only filtered out homozygous reference genotypes, accepting any variant found in both sequenced individuals. I rediscovered a missense variant in *ANK2* from a previous analysis through both SAMtools and BAMily. The gene is not known for playing a role in cardiogenesis, but has been associated with long-QT syndrome. However, the variant failed validation. Another particular missense variant stood out, being located in *CLTCL1*. This was again detected by both SAMtools and BAMily. 22q11.2 deletion syndrome typically encompasses *CLTCL1* among other genes, suggesting that it could play a role in some of the phenotypes observed in patients with the syndrome (Michaelovsky *et al.*, 2012). CHDs present in patients with 22q11.2 deletion syndrome can include IAA and VSD (Kobrynski and Sullivan, 2007). ExomeDepth did not reveal any CNV candidates.

In Family 5, I rediscovered three missense variants from one of my previous analyses. The variants were detected by both callers. The variants in *GHR*, *DMXL1* and *ZNG292* respectively, failed to validate. I also found new candidates, again with both callers. I found a missense variant in *HES1* predicted deleterious by MutationTaster, LRT and Polyphen2. *HES1* in mouse is expressed in the second heart field during cardiogenesis and is essential for outflow tract development (Rochais *et al.*, 2009). Loss of the gene in mice led to a displacement of the aorta over ventricles and a VSD (Rochais *et al.*, 2009). I also found another missense variant in *POPDC3* which, despite only being predicted to be deleterious by LRT was nonetheless of interest as it shares sequence similarity with *BVES* which is associated with another conotruncal defect, Tetralogy of Fallot; see chapter 5 (Wu *et al.*, 2013). Both these genes were discovered based on a chick gene ortholog which is preferentially expressed during chick cardiogenesis (Andree *et al.*, 2000). *HES1* and *POPDC3* are now awaiting validation. No candidate CNVs were found.

Given the difficulty in applying a carrier status to parents in Family 6, I ran BAMily twice for this family, each time with one of the parents identified as a disease carrier. Despite the large number of rare variants detected as a result, one candidate stood out due to the family's MAPCA phenotype. From the analysis in which the mother was predicted to be a disease-carrier, a missense variant in *NOTCH1* was found in the two siblings and the mother by both BAMily and SAMtools. The variant was predicted as deleterious by all pathogenicity predictors. *NOTCH1* is part of the Notch signalling pathway which plays a key role in cardiogenesis (Niessen and Karsan, 2008). MAPCA has been identified as a secondary cardiac anomaly in patients with ToF harbouring a *JAG1* mutation (McElhinney *et al.*, 2002). This is significant as *JAG1* is part of the Notch signalling pathway (Niessen and Karsan, 2008). In *NOTCH1*, mutations have been reported as causing a range of aortic valve diseases (Garg et al., 2005). No candidate CNVs were found.

Few variants were found by SAMtools and BAMily together in Family 7, none of which were convincing candidates. BAMily detected 51 rare variants predicted to be deleterious by MutationTaster, Polyphen2 and LRT. However, 30 were indels, most of which were likely false positives; which I determined using IGV. Multiple variants were found in the same genes. For example, 9 missense variants were found in *CTBP2* suggesting these were false positives. No candidates were retained. Additionally, no large rare CNVs were found to be shared among them.

In Family 8 both lists of variants predicted to be deleterious shared 3 variants, which corresponded to the candidates selected. Three candidates, in *CLIP4*, *SLC5A6* and *SLIT3* respectively were validated. The variant in *CLIP4* was a frameshift deletion. The gene is expressed in vascular endothelial growth factor-induced embryoid bodies (Rebhan *et al.*, 1998). The other two candidates are missense variants predicted to be deleterious by all pathogenicity predictors. *SLC5A6* is a sodium-dependent multivitamin transporter (Rebhan *et al.*, 1998). It has been found to be expressed in the embryonic mouse heart at E14.5. *SLIT3* is expressed in cardiomyocyte-like progenitor cells (Rebhan *et al.*, 1998). It is also expressed in rat vascular smooth muscle cells. Mice with a disrupted *Slit3* gene show right ventricular hypertrophy (Eppig *et al.*, 2015). No candidate CNVs were added to this list.

3.5. Discussion

3.5.1. BAMily performance

I assessed five different variant callers: SAMtools, GATK, GATK+FamSeq, GATK+Polymutt and BAMily. By comparing genotype and sequence data, I estimated the sensitivity and specificity of each of these variant callers given sequence data from three to five individuals, with different degrees of relatedness. For each variant caller, I recorded the total running time required to call variants for five individuals. For GATK, crucial pre-calling steps such as base quality score recalibration were factored into running time measurements.

When comparing standalone variant callers; SAMTools, BAMily and GATK, the latter two presented the best balance between sensitivity and specificity overall. The variant callers that rely on GATK, FamSeq and Polymutt, would typically be expected to build on GATK's sensitivity and specificity advantage, providing a marked improvement in at least one metric without significantly affecting the other. However, my tests revealed a different picture. Figure 18 and Table 5 and Table 6 show a gain of sensitivity when FamSeq was applied to GATK output in most of the scenarios I tested, but at the expense of specificity. The GATK+FamSeq calling strategy produced the lowest specificity across all five callers tested. With Polymutt, the gains for the scenarios we have tested were even less clear. Polymutt provided a slight advantage in specificity overall, best illustrated in Figure 18, but in many cases, GATK specificity remained unchanged where Polymutt was applied. The overall sensitivity attributed to Polymutt was lower than GATK's, as is apparent in Table 5. For the scenarios shown in Table 6 sensitivity remained unchanged whether PolyMutt was applied or not, with only one exception. From these measurements, I conclude that, for the scenarios explored, Polymutt and FamSeq do not lead to an improvement of the results already produced by GATK.

With data from five individuals, BAMily and SAMtools were able to produce variant calls in less than 9 hours running on a single core. By contrast, GATK's genotyping step alone required a similar amount of time to run on four cores. To this running time must be added that of recalibration and realignment steps that preceded genotyping and that also required multiple cores in order to run smoothly. GATK's total core running time was largely in excess of what was observed for SAMtools and BAMily. Thus, SAMtools and BAMily present a substantial running time advantage over GATK.

While SAMtools and BAMily present running times that fall within the same range, BAMily had an overall higher sensitivity and specificity. Accounting for running time, sensitivity and specificity together, BAMily presents a discernible advantage over GATK as well. While GATK and BAMily have comparable sensitivities and specificities across a number of scenarios, BAMily is able to deliver results in substantially less time. With my approach thus provides an advantageous balance between sensitivity, specificity and running time. Studies that aim to uncover variants present in one set of individuals and absent in another can greatly benefit from variant calling following this approach.

3.5.2. BAMily's usage in the analysis of 8 families

The design and implementation of a new approach to variant calling provided an opportunity to re-analyse families for which samples were sequenced, but no disease-causing candidates were found. Having undergone several rounds of analysis with no success, many of these cases are likely to remain unresolved with current WES approaches. Gilissen *et al.* (2012) reported a similar experience, with 60% of their WES projects centred on Mendelian disease leading to the discovery of causal variants, leaving many cases unresolved. However, some families could be elucidated with the shifts in approach that a variant caller such as BAMily can provide.

Using BAMily and SAMtools to analyse these 8 families revealed candidates in 5 out of 8 families. At the time of writing, many validations are still in progress (performed by laboratory colleagues and not part of my thesis work). However, for two families, candidates have already been succesfully validated. In Family 1, this is the case for a variant in *BLK*. The gene in which the mutation was found has been implicated in the development of KWE, which shares characteristics with HSP (Appel *et al.*, 2002). As mentioned briefly in Section 3.4.5, a candidate in *FAM111B* was found through a cross-collaboration predating my analysis. This variant had been missed from several analyses as it was classified by MutationTaster as benign. The presence of two other variants in the same gene in four distinct families with HSP provided evidence that the gene was probably misclassified (Mercier *et al.*, 2013). Although this variant appears to be the likeliest cause of HSP in Family 1, there is still room to consider the variant in *BLK* as a candidate or a contributing factor to the disease. In Family 8, three variants were validated, in *SLIT3*, *CLIP4* and *SLC5A6*. Of particular interest is the variant in *SLC5A6*; solute carrier family 5 member A6, as another gene encoding for a solute carrier, *SLC25A3*, has previously been implicated in cardiomyopathy (Mayr *et al.*, 2007). Demonstrating the disease-causing potential of any of the candidates listed in Table 7, through experimental assays for example, would provide further validation for the new approach to variant calling proposed implemented with BAMily.

3.6. Conclusion

In this chapter, I designed and programmed a new variant caller; BAMily, based on a new approach to variant calling. This approach focuses on detecting variants that are distributed in a specific way among a group of individuals. As such, it is primarily targeted at pedigrees. I evaluated the performance of the caller against other available programmes and confirmed it had comparable; and occasionally higher, accuracy coupled with its reduced computation demand. I incorporated the caller to a WES analysis of eight families for which previous analyses using older approaches had been unsuccessful. The use of the new caller resulted in the identification of strong candidate genes in two families that will require biological validation in future work. This completely new approach to variant calling will also be useful for future WES studies.

Chapter 4. *De novo* mutations in transposition of the great arteries

4.1. Summary

Transposition of the great arteries (TGA), while being one of the more common and severe forms of CHD, still remains poorly understood. The absence of familial precedents for the disease (in particular, relatively low sibling recurrence risks) has led to the hypothesis explored in this chapter that the manifestation of TGA in these individuals is the result of single, rare, large effect, *de novo* mutations. The following study focuses on the identification of *de novo* variants via whole-exome sequencing and analysis of family trios and one quartet. With this study I attempt to identify genes that play a role in normal development of the great arteries and that, when mutated, predispose to TGA. Three different variant callers are used towards fulfilling these aims: SAMtools, GATK and BAMily, with the latter acting as a filter. The variants called as a result are the main focus of the study.

4.2. Introduction

4.2.1. Transposition of the great arteries.

As briefly described in chapter 1, TGA is one of the most common and severe forms of CHD, but its origins remain poorly understood (Unolt *et al.*, 2013). With an incidence of 0.2 per 1,000 live births, it is the fourth most common CHD (Unolt *et al.*, 2013; Saremi, 2014). Untreated, TGA carries a 95% mortality rate in the first year of life (Saremi, 2014). Isolated cases represent half of all TGA while cases accompanied by extra cardiac defect account for 10% of all TGA (Martins and Castela, 2008). TGA takes on a number of forms, but can be broadly categorised as either dextro-transposition of the great arteries (D-TGA) or levo-transposition of the great arteries (L-TGA).

D-TGA is the most frequent form of TGA. Patients with D-TGA have switched arteries arising from the ventricles (Saremi, 2014). The aorta arises from the right ventricle, while the pulmonary artery arises from the left ventricle. This creates two separate circuits, with oxygen-poor blood circulating around the body while oxygen-rich blood is continually recirculated to the lungs. In a little over half of D-TGA cases, the pulmonary artery and aorta run parallel to each other instead of crossing as they would in a normal heart (Saremi, 2014).

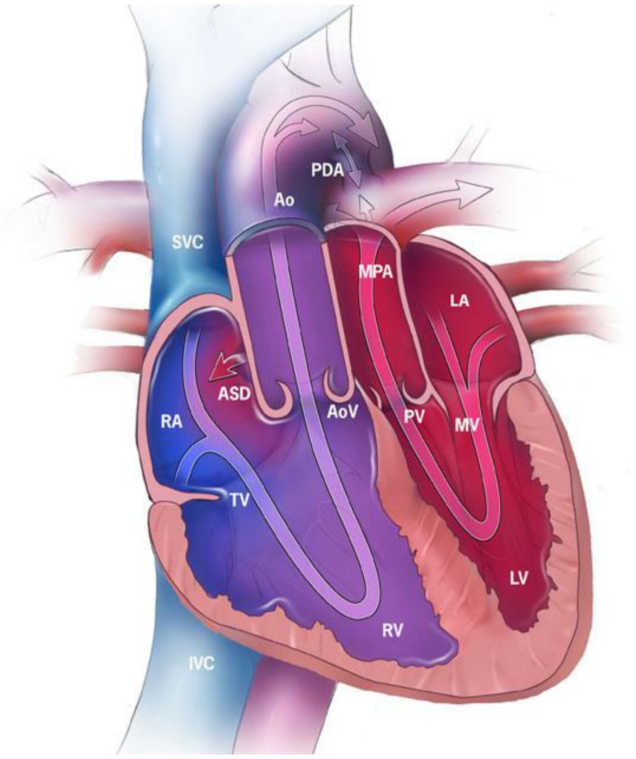

Figure 20. Dextro-transposition of the great arteries (D-TGA). In the above case, the TGA is accompanied by an ASD and PDA. Arrows indicate blood flow. Abbreviations: RA: right atrium, RV: right ventricle, LA: left atrium, LV: left ventricle, SVC: superior vena cava, IVC: inferior vena cava, MPA: main pulmonary artery, Ao: aorta, TV: tricuspid valve, MV: mitral valve, PV: pulmonary valve, AoV: aortic valve. ASD: atrial septal defect, PDA: patent ductus arteriosus. Image from the Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities available at: http://www.cdc.gov/ncbddd/heartdefects/TGA.html

L-TGA is the rarer form of TGA, accounting for less than 1% of all CHDs (Warnes, 2006). Patients with L-TGA have both switched arteries and atria. As a consequence, normal blood flow is maintained (Warnes, 2006). For this reason, L-TGA is also commonly referred to as congenitally-corrected TGA (C-TGA).

The exact mechanisms at work during embryological development that lead to TGA remain largely unknown (Unolt *et al.*, 2013). Two theories have been advanced on the subject. One theory, proposed by Goor and Edwards (1973), describes TGA as the result of an incomplete rotation of the aorta around the body's vertical axis, towards the left ventricle. The arteries find themselves aligned over the wrong ventricular outflow tracts. This theory is inspired by Goor's own observations on human embryos (Goor *et al.*, 1972). Goor and Edwards' theory also accounts for various other conotruncal defects, such as double outlet right ventricle (DORV) and tetralogy of Fallot (ToF). These defects also present a displacement of the arteries relative to the ventricles. In patients with DORV, the aorta and pulmonary artery arise from the right ventricle (Obler *et al.*, 2008). In patients with ToF, described in chapter 1 and chapter 5, the aorta is displaced over both ventricles while the pulmonary artery is stenotic (Nelson *et al.*, 2014). According to this theory, each distinct defect is the product of the same incomplete rotation, arrested at different stages of development, with TGA representing the most extreme end of the spectrum (Unolt *et al.*, 2013). Another theory, proposed by De la Cruz *et al.* (1977) sees TGA as the result of an abnormal development of the wall that separates the developing aorta and pulmonary artery. Instead of developing as a normal spiral, it develops linearly, the future aorta positioning itself above the right ventricle, leading to TGA. A recent study by Bajolle *et al.* (2006) lends more credibility to the first theory. By studying mice with mutated *Splotch* and *Pitxc2δc* genes, the authors were able to experimentally associate conotruncal defects with the arrested rotation of the aortic outflow tract.

A number of environmental risk factors are thought to contribute to the development of TGA. Martins and Castela (2008) point to maternal exposure to a number of teratogens; rodenticides, herbicides and antiepileptic drugs, and maternal diabetes as some risk factors. Maternal diabetes, has been identified as a risk factor for CHD in general (Wren *et al.*, 2003). Wren *et al*. (2003) report a 5-fold increase of CHD in newborns with diabetic mothers. However, maternal diabetes is especially relevant to TGA. TGA represents 14.4% of all CHD in newborns with diabetic mothers (Wren *et al.*, 2003). In this case, Wren *et al*. (2003) report a 17-fold increase in TGA in newborns with diabetic mothers (Wren *et al.*, 2003). Other environmental risk factors associated with TGA include maternal infection, ingestion of ibuprofen and exposure to ionising radiation during pregnancy and *in vitro* fertilisation (Unolt *et al.*, 2013).

The genetic underpinnings of TGA remain largely unresolved. Unlike many other CHDs, TGA is rarely observed in association with a chromosomal defect (Fahed *et al.*, 2013). Because of the extremely high and early mortality rate associated with the disease preceding recent surgical advances (Saremi, 2014), variants that contribute to disease risk are unlikely to be common in the population. For common variants to influence TGA risk would require some form of balancing selection given the deleteriousness of the disease. Additionally, highly penetrant dominant variants are unlikely to be passed on. Familial cases of the disease are thus expected to be particularly rare.

In the study by Burn *et al*. (1998) first mentioned in chapter 1, no cases of parent-offspring TGA were reported. Burn *et al*. (1998) suggest therefore that TGA is a sporadic defect. However, in their comment on the study, Digilio *et al*. (1998) point to several of their studies that support Mendelian inheritance of TGA in a select number of cases. In one such study, Digilio *et al*. (2001) found recurrence risks of CHD of 0.5% and 1.8% in parents and siblings of patients with TGA. The authors propose that some of this risk can be explained as cases of monogenic or multigenic inheritance. These few possible examples of familial TGA establish some genetic basis for the disease. However, the majority of TGA cases are sporadic (Digilio *et al.*, 2001). Given these facts, it can be hypothesised that *de novo* mutation may play a substantial role in accounting for sporadic TGA.

TGA is rarely accompanied by extracardiac defects, with the exception of laterality defects (Unolt *et al.*, 2013). TGA has been reported in cases of heterotaxy, a disease where organs develop on the opposite side of the body leading to a host of morphological defects (De Luca *et al.*, 2010). TGA has been frequently reported as part of asplenia syndrome, a type of heterotaxy characterised by the absence of a spleen (Unolt *et al.*, 2013). In mice, knockouts of *Smad2* and *Nodal* genes lead to TGA, with an added right pulmonary isomerism; where the morphology of the left lung takes on characters of the right lung, usually in more than half of cases (Unolt *et al.*, 2013). The association between laterality defects and TGA has led researchers to study genes involved in establishing the left-right body plan in isolated TGA cases (Goldmuntz *et al.*, 2002; De Luca *et al.*, 2010). A study by De Luca et *al.* (2010) found mutations in *FOXH1*, *ZIC3*, *NKX2.5* and *NODAL* in familial TGA cases. *CFC1* has also been implicated in patients exhibiting TGA (Goldmuntz *et al.*, 2002). A recent study of 362 cases of severe CHD included 47 patients with TGA (Zaidi *et al.*, 2013). Three de novo variants were uncovered in TGA patients: one in a known laterality gene, *SMAD2*, and two in genes not previously-associated with TGA or laterality defects, *NAA15* and *RAB10*. The WES study presented in this chapter is focused entirely on TGA. The probands used for the study comprise 32 patients exhibiting the disease. There is no previous family history of CHD in any of these cases. This leads to the hypothesis that the disease phenotype is brought about by single rare large effect *de novo* mutations.

4.2.2. Family-based WES for uncovering de novo mutations

Each individual is born with several variants that they do not share with either of their parents. Such mutations arise in the parental gametes, and occasionally during early embryonic development, and are referred to as *de novo* (Ku *et al.*, 2012). Before the advent of NGS, estimates of the rate of *de novo* per base mutation in a single genome for a single generation varied between 1.1×10^{-8} and 3×10^{-8} (Conrad *et al.*, 2011). Using WGS, Conrad *et al.* (2011) provide a similar, yet more accurate, per base estimate of 1.18×10^{-8}. This translates to an average of 74 new single nucleotide variants (SNVs) per newborn, with variation in rate between individuals found to be correlated with paternal age (Kong *et al.*, 2012; Veltman and Brunner, 2012b). With a genome that spans over 3 billion base pairs, any two unrelated individuals are exceedingly unlikely to share a *de novo* variant. *De novo* mutations will also tend to be more deleterious on average, having not been subjected to selection over multiple generations (Veltman and Brunner, 2012b). This leads Veltman and Brunner (2012b) to conclude that *de novo* mutations are the likely source of many sporadic disease cases.

An estimated 85% of all deleterious mutations are expected to lie in the protein-coding regions contained in the human exome, which itself covers 1% of the genome (Majewski *et al.*, 2011). Using the figures above, we can estimate the rate of *de novo* variants in the exome to be 0.74 on average per newborn. An earlier estimate by Robinson (2010) placed that figure at around 0.89. However, these mutations are expected to play a significant role in non-inherited; or sporadic, disease (Veltman and Brunner, 2012b). D*e novo* variants are more deleterious on average than inherited variants, that have been subject to natural selection over many generations (Veltman and Brunner, 2012b). A single *de novo* could therefore be responsible for a sporadic disease in an individual.

Both *de novo* CNVs and SNVs contribute to the risk of a number of neurodevelopmental diseases, such as autism spectrum disorder (AD) and schizophrenia. In chapter 1, I described one study by Sanders *et al*. (2012) which revealed an excess of non-synonymous *de novo* variants in 238 patients with autism spectrum disorder (AD) when compared with 200 healthy siblings. Focusing on truncating *de novo* variants in genes expressed in the brain revealed a large contrast, with an odds ratio of truncating to synonymous variant in cases and controls of 5.65. A more recent example, a large-scale study of *de novo* variants in 1,078 AD patients compared with 343 unaffected siblings, revealed an excess of genes with more than two *de novo* non-synonymous variants in patients (Samocha *et al.*, 2014). Samocha *et al*. (2014) compare their results against a model of the rate of *de novo* mutation per gene; with the rate of mutation overall and in functional subsets factored in as well, for each variant type. The number of mutation events in healthy siblings was found to fit the model. However, an excess of truncating variants was found in AD patients. Furthermore, an excess of genes with at least 2 *de novo* mutations was found in AD patients for all variant types except synonymous variants. Here again, results from the healthy siblings fit the model's expectations. These findings strongly implicate *de novo* mutation affecting specific protein-coding genes in the development of sporadic AD. One study of AD by O'Roark *et al*. (2012b) was able to establish a paternal bias in *de novo* variants, accounting for the increase of AD risk with paternal age using a combination of gene sequencing. O'Roark *et al*. (2012b) sequenced the exome of 209 trios and 50 unaffected siblings, using markers for haplotype phasing to determine the origin of *de novo* variants. Studies suggest that *de novo* variants could also play a role in the manifestation of schizophrenia (Xu *et al.*, 2011). While schizophrenia shows a strong familial preponderance, patients with no family history have been reported (Xu *et al.*, 2011). Several studies of schizophrenia have focused on the contribution of *de novo* copy number variants (CNV) to the disorder, including one study by Xu *et al*. (2008) which found that *de novo* CNVs were 8 times more frequent in 152 schizophrenia patients then in 159 unaffected patients. To identify *de novo* SNVs potentially contributing to sporadic cases of schizophrenia, Xu *et al*. (2011) sequenced the exome of 53 parents-offspring trios; in which only the offspring had schizophrenia, and 22 unaffected trios. Xu *et al*. (2011) found a large excess of non-synonymous *de novo* variants in patients with schizophrenia compared with healthy controls. They also found a ratio of *de novo* non-synonymous variants to *de novo* synonymous variants in the case population exceeding what would be expected in the general population. It remains difficult to demonstrate which particular *de novo* variants contribute to these neurodevelopmental disorders, particularly if the genes have not previously been implicated in the disease. However, in all these studies, the excess of non-synonymous variants; especially truncating variants, strongly suggest an important role for *de novo* variants in these pathologies.

For CHD, the role of *de novo* mutation has been investigated in a study conducted by Zaidi *et al.* (2013). The authors have chosen to focus on severe sporadic CHD cases; which include TGA cases. Presumably, this choice was operated under the assumption that severe cases would be more likely to be caused by *de novo* variants. The authors selected 4,169 genes known to be highly expressed during mouse cardiogenesis and compared the incidence of *de novo* variants in gene orthologs in 362 severe CHD patients and 264 controls originating from parent-offspring trios. Zaidi *et al.* (2013) uncovered a significant excess of *de novo* changes; both synonymous and non-synonymous, in the selected genes in CHD cases compared to controls. The contrast between cases and controls was more pronounced for truncating variants, amounting to an odds ratio of 7.5. This excess was particularly pronounced in histone-modifying genes. Many of the genes in which *de novo* variants were found involved the writing, removing and reading of histone methylations. This class of genes is known for its regulation of gene expression (Bruneau and Srivastava, 2014). Mutations in those genes could therefore have far-reaching consequences on cardiac cell differentiation, potentially leading to many of the defects observed in CHD patients (Bruneau and Srivastava, 2014). Based on the rate of non-synonymous *de novo* variants in these gene orthologs in cases compared to controls, Zaidi *et al.* (2013) put forward the idea that *de novo* mutation contributes 10% of severe CHD (95% confidence interval: 5-15%). Ultimately, this estimate relied on similarities in high gene expression during cardiogenesis between mice and humans and could therefore be an underestimate (Zaidi *et al.*, 2013). It is also important to emphasise here that this conclusion pertains only to severe CHD. In a commentary, Bruneau and Srivastava (2014) reported this last conclusion by Zaidi *et al.* (2013) as applicable to sporadic CHD in general. However, it is possible that *de novo* mutation does not play as great a role in milder forms of sporadic CHD.

The study of putative *de novo* mutations underlying genetic diseases requires the sequencing of disease patients and their unaffected parents (Bamshad *et al.*, 2011). For rare genetic diseases, it is possible to first identify potentially *de novo* candidate mutations in cases via WES and then confirm which mutations are *de novo* via gene-targeted Sanger sequencing of the entire family trio or by using modified molecular inversion probes (O'Roak *et al.*, 2012a). An early study by Hoischen *et al.* (2010) illustrates this practice with the WES of four patients with Schinzel-Giedion syndrome. Patients with Schinzel-Giedion syndrome present severe mental retardation, distinctive facial features and several organ and bone abnormalities (2010). In each of the four patients, Hoischen *et al.* (2010) uncovered a variant at a different locus in *SETBP1*. Confirmation of the variant and its *de novo* nature was obtained through targeted sequencing of the corresponding family trios. Hoischen et al. (2010) also used targeted sequencing on eight more cases, revealing additional variants, the total set clustering within an 11bp exonic region. However, locus heterogeneity is common in genetic disease, whereby mutations occurring in distinct genes lead to the same disease trait (Veltman and Brunner, 2012b). In this context, the putative disease-causing *de novo* mutations in individuals exhibiting a shared disease phenotype could be located in different genes, making it difficult to pick out likely candidate genes. Studies involving a disease trait for which locus heterogeneity is suspected call for entire trios to be analysed via WES. Neurodevelopmental disorders illustrate the fact that *de novo* mutations at different loci can lead to the same trait (Hoischen *et al.*, 2014). Hoischen *et al*. (2014) report that schizophrenia could be the product of mutation in more than 500 distinct genes. The first study involving the WES of parent-offspring trios was performed by Vissers *et al.* (2010) and centred around 10 patients with sporadic mental retardation. *De novo* non-synonymous variants were identified in 9 genes, with six variants retained as probably disease-causing based on functional evidence. Crucially, these six *de novo* variants were all found in distinct genes. The CHD study by Zaidi *et al.* (2013) previously mentioned provides another example of locus heterogeneity. A total of 37 genes in cases were found to harbour potentially disease-causing *de novo* variants, with 34 genes harbouring only one variant. Furthermore, the authors estimate to total number of disease-associated genes to be around 400.

The following study involves 32 patients exhibiting TGA. By sequencing the 31 corresponding parent-offspring trios as well as a single nuclear family of four (the additional sibling being unaffected), I seek to uncover *de novo* mutations potentially responsible for triggering TGA. Multiple variant callers are enlisted for this purpose, producing a consensus list of candidate variants. For those variants most likely to be pathogenic, validation work was performed by colleagues based in the laboratory. Several layers of annotation, data-mining and pathway analysis are used to elucidate a potential association between each variant and TGA.

4.3. Methods

The sequencing of the samples required for this study was shared between three institutes: Newcastle's Institute of Genetic Medicine (IGM) and the Glasgow Polyomics (GP) research facility in the United Kingdom and the The McGill University and Génome Québec Innovation Centre (MUGQIC). The share of sequencing was distributed as shown in Table 9. Target enrichment was performed using Agilent's SureSelect^{XT} Human All Exon 50Mb kit V4 (Agilent Technologies, 2015). The quality control applied to sequencing does not differ from what was described in Chapter 2.

Institute	Patients sequenced	families complete after sequencing	First sequencing run date	Last sequencing run date
IGM	43	12	23/08/2012	08/03/2013
GP	24	22	16/05/2013	25/07/2013
MUGQIC	30	32	05/03/2014	20/10/2014

Table 9. Distribution and duration of WES for institutes involved. Sequencing was spread across the duration of the PhD.

The sequencing of 32 families was spread across the entire PhD project. As such, I have re-analysed families on multiple occasions, prompting progressive refinements to the methods used. For the purpose of this thesis however, I will focus primarily on the methods used in my final analysis.

I performed sequence alignment for 97 samples using the aligner BWA v0.7.4 (Li and Durbin, 2009). Originally, I used NovoAlign v2.7.13 (Novocraft, 2014b). However, NovoAlign's main advantage, namely its higher sensitivity, comes at the price of a longer running time (Novocraft, 2014a). With the significant number of individuals involved in this study and limited computing resources at first, I eventually selected BWA v0.7.4 (Li and Durbin, 2009) over NovoAlign. For consistency, I re-aligned the few samples originally aligned using NovoAlign with BWA. I removed duplicates using Picard (The Broad Institute, 2015c).

During the project, I used three different variant callers to identify variants present in the aligned sequence samples:

- SAMtools V0.1.18 (Li *et al.*, 2009a)
- GATK UnifiedGenotyper v.2.2.9 (DePristo *et al.*, 2011)
- BAMily (For more details, see chapter 3)

SAMtools and GATK UnifiedGenotyper were used to create a consensus list of high-confidence variants. There are some discrepancies between variant callers, even when only high-confidence variants are considered (O'Rawe *et al.*, 2013). These discrepancies are particularly evident with indels which are generally more difficult to detect than SNVs (O'Rawe *et al.*, 2013). The assumption here is that variants reported by two widely-used callers are more likely to reflect the existence of actual variants. Consensus between callers therefore serves as a final quality filtering criterion for candidate variant selection. BAMily's strength resides in its ability to detect variants present across multiple samples and absent across others, even if the evidence for occurrence in single samples is weak. Using BAMily to detect variants present in a single proband will produce many more false positives than in SAMtools and GATK. However, BAMily can help with the exclusion of false positives that arise from variants being called incorrectly in parents. In this study, BAMily therefore plays the role of a filter. The process by which BAMily calls variants across multiple samples is described in chapter 3. The consensus between SAMtools and GATK, with the filtering step provided by BAMily, results in a final consensus list of variants.

I performed variant calling using SAMtools and BAMily by family unit. For each variant site, SAMtools assigns a genotype to each member of the trio. I searched variant sites at which the genotypes were consistent with a *de novo* variant in the offspring. In other words, I selected variant sites where the offspring was predicted to be heterozygous for the variant while both parents were predicted as homozygous for another base, most likely the reference. BAMily does not require this filtering step, as the expected variant assignment for each individual must be established prior to variant calling, as discussed in chapter 3. I therefore set BAMily to provide calls for variants present in the offspring, but absent in parents; and sibling in the quartet. With GATK's UnifiedGenotyper, all aligned samples were called for variants in a single run. As detailed in chapter 2, variant calling with the UnifiedGenotyper is preceded by a number of realignment and recalibration steps. As with SAMtools, I filtered for variant sites for which the genotype assignment suggested a *de novo* variant. For all variant callers, variants outside the exome target regions were excluded. As described in chapter 2, I only considered variants that passed the high-confidence quality threshold Q30. As variants found to be *de novo* are unlikely to reoccur in any of the other trios, I excluded variants found in any of the 32 pairs of parents and the unaffected offspring.

The resulting variants were then annotated using ANNOVAR (Wang *et al.*, 2010). Variants were retained if they represented non-synonymous changes in an exonic or splice region and were rare. A variant was considered rare if it had a minor allele frequency (MAF) of <1% according to both the 1000g (Abecasis *et al.*, 2012) and the NHLBI Exome Sequencing Project (ESP) (NHLBI, 2015) population databases. In addition to the population databases from ANNOVAR (Wang *et al.*, 2010), an in-house list of variants detected in previous exome studies; totalling 418 exome sequences, helped to exclude more variants from the list.

The filtered list of variants detected by SAMtools and GATK produced several candidates. Candidates were further prioritized depending on the consensus of multiple functional prediction tools. Originally, these were limited to MutationTaster (Schwarz *et al.*, 2010) and Polyphen2 (Adzhubei *et al.*, 2010), but LRT (Chun and Fay, 2009) and MutationAssessor (Reva *et al.*, 2011) were eventually incorporated into the analysis. The full list of prioritised variants is presented in the Section 4.4.

The Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013) provided an additional tool for inspecting candidates and eliminating variants in high repeat regions or indels that appeared to be artefacts when viewed in their full sequencing context. This was followed by a review of the genes candidate variants appeared in, using the Online Mendelian Inheritance in Man (NCBI, 2015) database as well as several resources housed by the GeneCards compendium (Rebhan *et al.*, 1998). Variants selected as candidates were communicated to colleagues (see acknowledgments) at the IGM and at Manchester University's Institute of Cardiovascular Sciences for validation through Sanger sequencing.

I investigated the genes in which *de novo* variants were found using a number of pathway analysis tools. I first analysed genes in STRING v.10 (Jensen *et al.*, 2009), followed by EnrichNet (Glaab *et al.*, 2012). STRING provides a method for evaluating and visualising potential functional links between proteins (Jensen *et al.*, 2009). The functional links are built around several categories of evidence; represented as color-coded edges in STRING's network view, that tie the activity of two proteins; or their putative homologs in other species, together (Jensen *et al.*, 2009). Evidence in each category is accompanied by a confidence score which adds up to a combined score; on a scale from 0 to 1. Functional links with a score above 0.4 are represented as edges in STRING's network view (Jensen *et al.*, 2009). I looked at the functional links between the protein products of the genes from my final list of variants. Additionally, I looked for enrichment in gene ontology (GO) terms (Gene Ontology, 2004) and KEGG interaction pathways (Kanehisa and Goto, 2000). Applying my own scripts to STRING data, I determined the shortest functional pathways between protein products. Using EnrichNet, I interrogated a number of pathway databases. I also submitted the genes in which *de novo* variants were found in SAMtools and GATK separately.

Recently, the ExAC (The Broad Institute, 2015a) population database was released for public use. The database contains variant calls from 60,706 unrelated individuals. I therefore used this database to further annotate my final list of variants.

This study has undergone a number of iterations over the course of three years. The following section presents the results of the WES study in its final iteration.

4.4. Results

Identifying potentially disease-causing *de novo* variants required filtering large amounts of variant data. Within exome targets, SAMtools detected an average of 38,137.4 high-confidence variants (Q30) per individual while GATK presented an average of 45,690.6 high-confidence variants per individual. These large quantities of variant data required various rounds of filtering to be applied in order to restrict each list to variants most likely to result from actual *de novo* mutations with disease-causing potential. Table 10 provides an average for all of these steps for the probands.

Variant caller	Total	Detected as *de novo*	Rare	In exon or splice site	Non-synonymous	Not in-house*	Detected by BAMily
SAMtools	38,656.3	21.7	15.1	6.6	4.8	3.6	1.2
GATK	45,525.6	27.9	12.6	7.1	4.9	3.4	1.2

Table 10. Average number of high-confidence variants in probands at different stages of the filtering process by variant caller. *Not previously detected in-house in 418 exomes.

Consistent with the extremely low number of *de novo* variants expected in a single exome; estimated to be 0.74 as described in Section 4.2.2, the number of high-confidence variants predicted as *de novo* by both SAMtools and GATK for any proband was many orders of magnitude lower than its total number of high-confidence variants. However, to get to a final list of *de novo* variants with fewer false positives, a number of steps were required. *De novo* variants are no more likely to occur at a common variant site than anywhere else in the genome. Variants that coincided with common variants were thus almost certainly false positives. Excluding non-rare variants, as well as any variant that was not in a splice site or exon, brought the average number of SNVs and indels per individuals closer to the estimate of Section 4.2.2.

The number of predicted non-synonymous *de novo* SNVs and indels called for each proband varied although tended to converge with each filtering step. The average number of variants per proband, the range of value across probands and the total number of variants by variant caller are given in Table 11. The average number of variants identified by SAMtools and GATK was only slightly in excess of the expected average. The consensus between the two callers is likely to discard some of the remaining false positives with BAMily providing an additional filter for false positives.

Variant Caller	**Average (per proband)**		**Range**		**Total**	
	SNVs	indels	SNVs	indels	SNVs	Indels
SAMtools	2.3	1.3	0-9	0-4	73	41
GATK	3.3	0.1	0-20	0-1	105	4

Table 11. Average and range of non-synonymous de novo variants per proband once all filtering has been applied. Also given is the total number of variants by variant caller.

Once filtering with BAMily was applied, the number of variants produced by each caller and both caller were further reduced to the numbers shown in Figure 21.

4.4.1. De novo variants identified by consensus.

Of the 47 variants found post-filtering, around 51% were shared by both SAMtools and GATK UnifiedGenotyper. This represented a total of 23 SNVs and a single indel. GATK, with its IndelRealigner step, provided a conservative number of indel detections. Before the application of the filter provided by BAMily, GATK had detected a total of 4 indels, 3 of which were shared with SAMtools. By contrast, SAMtools produced a list of 42 indels, most of which were likely to have been artefacts. The share of SNVs and indels identified by SAMtools and GATK post-filtering; including the application of GATK, is shown in Figure 21.

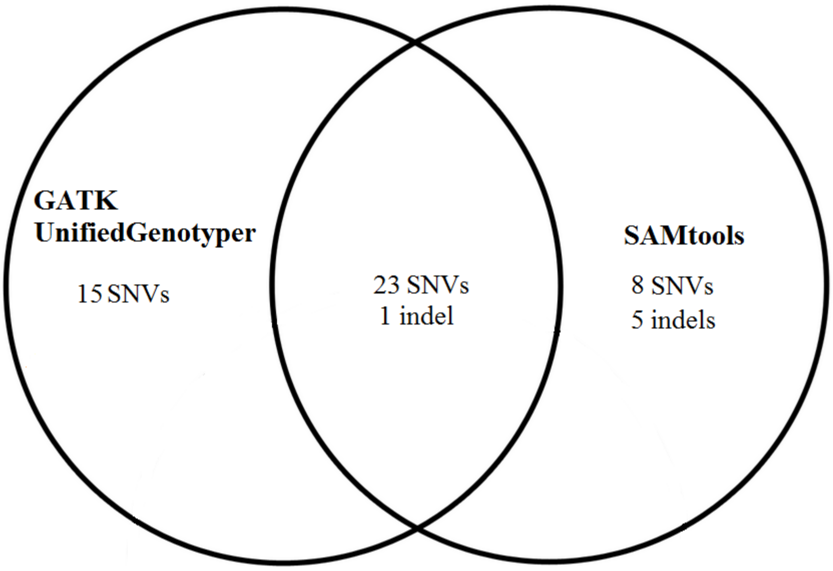

Figure 21. Distribution of variants detected in 32 probands post-filtering in both variant callers. BAMily was used as a filter.

Variant were found among 14 out of a total of 32 probands. In total, 24 variants were found, corresponding to a ratio of *de novo* variants per proband of 0.75, close to the 0.74 described in Section 4.2.2. The variants, and the genes and families in which they occurred, are summarised in Table 12.

The 24 variants found post-filtering included 2 nonsense variants and a single frameshift deletion. These were found in genes *ZNF227*, *GDPGP1* and *ENTPD2* respectively. The possible pathogenicity of the remaining 21 missense variants was based on pathogenicity predictions from four functional predictors as shown in Table 13. 16 out of 21 missense variants were flagged as probably pathogenic by at least one of the functional predictors. Among these 16 variants, 4 variants were predicted as at least possibly pathogenic; or equivalent, by Polyphen2, LRT and MutationTaster, with support from MutationAssessor. The 4 variants were found in *ANGPTL2*, *PROK1*, *PHLPP2* and *PDE4D* respectively. Each variant in the consensus list occurred in a distinct gene. I will henceforth refer to each variant by the gene in which it occurred.

The genes in which the variants in the consensus list occur have not previously been associated with TGA or other conotruncal defects. None have been identified as laterality genes. None of these genes are orthologs for the 276 genes highly expressed in the developing mouse heart as compiled by Zaidi *et al*. (2013). The closest association to this list is the presence of a paralog of *KCNJ12*, known as *KCNJ2*.

Family	Gene	Position (hg19)	Nucleotide and amino acid change (transcript)
33	*ANGPTL2*	chr9:129856079	c.G944C:p.G315A (ENST00000373425)
	KCNJ12	chr17:21319408	c.G754A:p.D252N (ENST00000331718, ENST00000583088)
	LMOD3	chr3:69168173	c.G1333A:p.E445K (ENST00000420581, ENST00000489031)
	ZNF419	chr19:58004899	c.G974A:p.S325N (ENST00000221735), c.G977A:p.S326N (ENST00000424930)
70	*CPT2*	chr1:53675809	c.A463G:p.M155V (ENST00000371486)
85	*ZNF577*	chr19:52376725	c.T518C:p.L173P (ENST00000301399)
216	*RBP5*	chr12:7280940	c.A148G:p.M50V (ENST00000266560)
222	*FAM208B*	chr10:5789201	c.C3817A:p.L1273I (ENST00000328090)
	PROK1	chr1:110998906	c.T251C:p.L84P (ENST00000271331)
245	*DCDC1*	chr11:31327850	c.A520G:p.I174V (ENST00000452803)
	GREB1	chr2:11758998	c.G991A:p.V331M (ENST00000396123)
	RAD52	chr12:1034628	c.G531C:p.K177N (ENST00000358495)
280	*PHLPP2*	chr16:71713341	c.T202A:p.S68T (ENST00000568004), c.T988A:p.S330T (ENST00000393524), c.T988A:p.S330T (ENST00000568954)
	ZNF227	chr19:44732650	c.C112T:p.R38X (ENST00000313040)
311	*GDPGP1*	chr15:90785068	c.C928T:p.R310X (ENST00000558017) *
312	*PDE4D*	Chr5:58271514	c.C1110A:p.D370E (ENST00000317118), c.C1077A:p.D359E (ENST00000358923), c.C1983A:p.D661E (ENST00000340635), c.C1575A:p.D525E (ENST00000360047) ,c.C1617A:p.D539E (ENST00000405755), c.C1593A:p.D531E (ENST00000503258), c.C1791A:p.D597E (ENST00000507116), c.C1800A:p.D600E (ENST00000502484)
	ENTPD2	chr9:139945759	c.450delC:p.Y150fs (ENST00000312665, ENST00000355097)
388	*COL11A2*	chr6:33134330	c.C4031T:p.A1344V (ENST00000361917) **, c.C4094T:p.A1365V (ENST00000374708) **, c.C4352T:p.A1451V (ENST00000341947) **
	GPR17	chr2:128409078	c.G853A:p.V285I (ENST00000393018, ENST00000272644)
478	*KRTAP4-6*	chr17:39296424	c.C316T:p.R106C (ENST00000345847)
501	*HECTD4*	chr12:112622770	c.G9484A:p.V3162M (ENST00000377560)**, c.G9562A:p.V3188M (ENST00000550722)**
503	*CREB3L3*	chr9:4157266	c.C431T:p.P144L (ENST00000078445, ENST00000602147, ENST00000602257), c.C428T:p.P143L (ENST00000595923)
	SNX13	chr7:17937012	c.T70C:p.F24L (ENST00000409604, ENST00000428135)
540	*ZFHX3*	chr16:72829270	c.T4569A:p.S1523R (ENST00000397992), c.T7311A:p.S2437R (ENST00000268489)

Table 12. The 24 variants from the consensus list identified by family. For brevity, only well supported transcripts are shown here unless unavailable. Well supported transcripts are described in Ensembl as being supported by one or more non-suspect mRNA. *Transcript has support level 2 meaning that the best supporting mRNA has been flagged as suspect **Transcript has support level 5 meaning that the reported structure is not constructed from any single transcript. More on transcript support levels available at: http://www.ensembl.org/Help/Glossary?id=492

Family	**Gene**	**pathogenicity prediction***			
		Polyphen2	LRT	MutationTaster	MutationAssessor
33	*ANGPTL2*	probably damaging	deleterious	disease causing	medium
	KCNJ12	benign	deleterious	disease causing	neutral
	LMOD3	benign	neutral	polymorphism	medium
	ZNF419	possibly damaging	N/A	polymorphism	low
70	*CPT2*	benign	neutral	disease-causing	neutral
85	*ZNF577*	probably damaging	N/A	polymorphism	high
216	*RBP5*	benign	neutral	disease-causing	low
222	*FAM208B*	benign	neutral	polymorphism	low
	PROK1	probably damaging	deleterious	disease-causing	medium
245	*DCDC1*	benign	neutral	polymorphism	neutral
	GREB1	benign	deleterious	polymorphism	neutral
	RAD52	benign	deleterious	disease-causing	medium
280	*PHLPP2*	probably damaging	deleterious	disease-causing	low
	ZNF227	Nonsense mutation			
311	*GDPGP1*	Nonsense mutation			
312	*PDE4D*	Probably damaging	deleterious	disease-causing	medium
	ENTPD2	Frameshift deletion			
388	*COL11A2*	benign	deleterious	polymorphism	neutral
	GPR17	benign	neutral	polymorphism	neutral
478	*KRTAP4-6*	N/A	unknown	N/A	medium
501	*HECTD4*	benign	N/A,	N/A	neutral
503	*CREB3L3*	benign	N/A	polymorphism	neutral
	SNX13	probably damaging	deleterious	N/A	low
540	*ZFHX3*	benign	deleterious	disease-causing	low

Table 13. Pathogenicity prediction of each variant based on the output of four functional predictors. *Possible predictions for each program are as follows: Polyphen2: probably damaging, possibly damaging, benign; LRT: deleterious, neutral, unknown; MutationTaster: disease-causing, disease-causing automatic [nonsense or known disease-causing in dbSNP], polymorphism, polymorphism automatic [any known variant in dbSNP not identified as disease-causing]; MutationAssessor: high, medium, low, neutral.

Submitting the 24 corresponding genes to STRING (Jensen *et al.*, 2009) revealed possible functional links between some of the protein products. Although the set of proteins was not enriched in interactions (*p*-value=1.63E-1), possible functional links were found between PDE4D, PHLPP2 and ZNF227. While an interaction between 3 proteins among 24 is not statistically significant (1.03E-1), the interactions were nonetheless noteworthy in that they occurred between three particularly strong candidate genes. The functional links, between these proteins and between GPR17 and PROK1, are represented by STRING's network view shown in Figure 22. Both functional links to PHLPP2 were predicted on the basis of several types of evidence, including the interaction of putative homologs in other species. The functional link between PHLPP2 and ZNF227 gave a combined score of 0.44. This was based on the co-mention of these two proteins; as well as putative homologs, in five publications; as determined through their abstract. There was also evidence of co-expression of homologs in *plasmodium falciparum* as well as evidence for protein-protein interactions in several species; specifically evidence of protein binding in *Drosophila melanogaster*. The functional link between PHLPP2 and PDE4D gave a combined score of 0.82. This was also based on co-mention in five publications; but only for putative homologs, and co-expression of homologs in *plasmodium falciparum*. Evidence of protein-protein interactions was based on homologs in *Saccharomyces cerevisiae.* The evidence for a functional link between GPR17 and PROK1 was particularly strong, leading to a combined score of 0.9. This score was based on both proteins appearing in the G alpha (q) signalling pathway stored in the Reactome Pathway database (Croft *et al.*, 2014). It is worth noting however that *GPR17* was predicted to be benign by a consensus of functional predictors. No significant GO term or KEGG pathway enrichments were found.

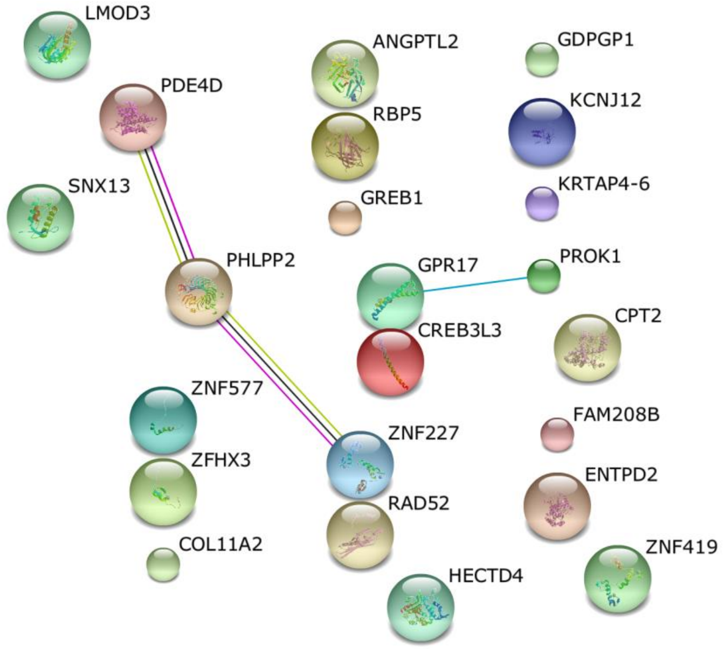

Figure 22. STRING network view of the protein products and predicted functional links for 23 genes in which *de novo* variants were found (DCDC1 was not available for network view). The different types of evidence for functional links between proteins are represented by edges. Evidence for functional links between proteins was derived from text-mining (yellow), experimental work (pink) and co-expression data (black) and association in a curated databases (blue).

Using the list of functional links between proteins in STRING, I determined the functional distance between each of the 24 proteins corresponding to my list of genes. The number of functional links separating each gene via its protein product is given in the Appendix, Figure S1. Ultimately, the analyses using STRING were inconclusive, providing no additional information by which the variants of interest could be further prioritised.

Submitting the 24 genes from the consensus list via EnrichNet (Glaab *et al.*, 2012) did not provide further evidence of enrichment for specific pathways. I submitted two additional gene lists to EnrichNet based on the list of *de novo* variants found in SAMtools and GATK separately. No pathway enrichment was found for either of those lists of genes.

DNA samples for 11 trios were sent away for Sanger validation work. The validations centred on 16 of the 24 *de novo* variants called by SAMtools and GATK, selected according to functional prediction; showed in Table 13. Variants were selected for validation work if they were considered as at least probably deleterious; or equivalent, by one of the functional predictors used. This meant that variants which were considered only possibly deleterious by one predictor; such as the variant in *ZNF419*, but otherwise benign, were not selected. Table 14 lists the gene variants for which validation work was requested and the subsequent results.

Family	**Gene**	**Confirmed *de novo*?**
33	*ANGPTL2*	yes
	KCNJ12	Parental traces unclear, gene is possibly paternally inherited
70	*CPT2*	yes
85	*ZNF577*	yes
216	*RBP5**	yes
222	*PROK1*	yes
245	*GREB1*	yes
	RAD52	yes
280	*PHLPP2**	yes
	*ZNF227**	yes
311	*GDPGP1*	yes
312	*PDE4D*	yes
	ENTPD2	yes
388	*COL11A2*	yes
503	*SNX13*	yes
540	*ZFHX3*	Validation in progress

Table 14. DNA samples from 11 families were sent away for validation work. A total of 14 variant sites were confirmed to harbour a *de novo* variant in the proband. Validation work for *KCNJ12* remains ambiguous,

variant potentially paternally inherited. *Validation work at these variant sites was done in an earlier iteration of the project, but the variants were rediscovered in this final iteration.

All but two of the variants detected by SAMtools and GATK were confirmed to be *de novo*; with one variant still awaiting validation. This proves that my pipeline is able to accurately identify variants that appear to be *de novo* from blood samples. It is still possible that some of these variants are not *de novo,* but the result of mosaicism, which I will explain in Section 4.5.

4.4.2. Evidence of processes relevant to cardiogenesis in genes for which de novo variants were uncovered using WES.

Available information on each of the 16 genes selected in the previous section was obtained using GeneCards (Rebhan *et al.*, 1998) and the OMIM (NCBI, 2015) compendiums. Here I provide a summary of findings in genes that support a potential role in the normal development of the heart.

One approach taken in uncovering a role for the genes in cardiogenesis was to look at whether the mRNA product of that gene had been reported as differentially expressed in embryonic tissues relating to the cardiovascular system. Results for 3 genes are shown in Table 15. These results rely largely on the LifeMaps database, which is part of the GeneCards suite (Rebhan *et al.*, 1998). It is worth noting that, for a number of genes, this information was not available. Therefore, the absence of reported expression in the embryonic heart tissue cannot be interpreted as an absence of a role in cardiogenesis.

Gene	**Expression in embryonic heart tissue**
ANGPTL2	Atrioventricular canal cells in the dorsal aorta and outflow tract
GREB1	atrioventricular canal cells
ENTPD2	atrioventricular node cells

Table 15. Positively differentiated expression in embryonic heart tissue of mRNA reported for genes of interest. Data collected by LifeMaps, part of the GeneCards suite (Rebhan et al., 1998).

Expression in adult heart can also provide some clues as to a potential role in cardiogenesis. *ANGPTL2* has been reported as highly expressed in the adult heart (Rebhan *et al.*, 1998; Kim *et al.*, 1999). *ENTPD2* was also reported as expressed in the adult heart (Rebhan *et al.*, 1998). One of *PDE4D*’s many protein products, isoform 7, has been detected in heart and skeletal muscle (Rebhan *et al.*, 1998). Once again, it is difficult to conclude that a gene was not expressed in heart tissue where annotations were not present or potentially incomplete. The expression of *PROK1* was identified as limited to endocrine tissue in the ovary, testis, placenta and adrenal gland, which could be an argument for ruling it out as a contributor to cardiogenesis (LeCouter et al., 2003). Nonetheless, there is no information on the gene's expression in embryonic heart (Rebhan *et al.*, 1998).

Using the Mouse Genome Informatics (MGI) database (Eppig *et al.*, 2015), I surveyed mouse knockouts in genes homologous to the genes of interest. Once again, I looked for genes which, when mutated, lead to an abnormal cardiovascular phenotype or embryonic lethality. The results are shown in Table 16.

Gene	**Mouse gene homolog**	**Abnormal cardiovascular phenotype in mouse**
ANGPTL2	Angptl2	None
COL11A2	Col11a2	None
CPT2	Cpt2	None
ENTPD2	Entpd2	None
GDPGP1	Gdpgp1	None
GREB1	Greb1	None
KCNJ12	Kcnj12	None
PDE4D	Pde4d	Increased cardiac muscle contractility
PHLPP2	Phlpp2	None
PROK1	Prok1	None
RAD52	Rad52	None
RBP5	Crabp1	None
SNX13	Snx13	Abnormal cephalic vascularization, small number of capillaries in neural folds (embryonic lethality during organogenesis).
ZNF227	None	/
ZNF577	None	/
ZNFHX3	None	/

Table 16. Abnormal cardiovascular system phenotype in mutated mice. Where mice homologs of the genes were found, gene knockouts were available. Data collected by the Mouse Genome Informatics database (Eppig *et al.*, 2015**).**

Mutated mice in two homologs were found to lead to an abnormal cardiovascular phenotype. However, neither led to TGA or any other conontruncal defect in mice.

4.4.3. Re-evaluation of the data using ExAC.

The recent release of ExAC (The Broad Institute, 2015a) presents an additional opportunity to establish whether the variants in the consensus list have been previously reported. In Table 17 are shown the allele counts in ExAC; with corresponding MAF, for each variant found both in the consensus list and in the ExAC database. This list includes variants that were confirmed to be *de novo* through Sanger sequencing, such as *GDPGP1* and *GREB1*. The variant occurring in *KCNJ12*, which could not be confirmed as *de novo*, was found in ExAC with a MAF which suggest that the variant is unlikely to be *de novo* in the proband.

Gene	**Allele count**	**MAF**
ANGPTL2	0	/
COL11A2	0	/
CPT2	0	/
ENTPD2	0	/
GDGP1	1	8.24E10^{-6}
GREB1	11	1.73E10^{-4}
KCNJ12	229	1.89E10^{-3}
PDE4D	0	/
PHLPP2	0	/
PROK1	0	/
RAD52	0	/
RBP5	0	/
SNX13	0	/
ZNF227	0	/
ZNF577	0	/
ZNFHX3	0	/

Table 17. Variants found by SAMtools and GATK that are also present in ExAC, listed by the gene these occur within.

Some of the cohorts included in ExAC belong to disease-specific population studies. Some of these studies involve heart disease patients, but these focus on the development of heart disease later in life rather than CHD. Given that CHD can accompany other diseases, a few patients with CHD phenotypes could have been included in one of these disease-specific populations. Therefore, it is difficult to determine what the presence of some of the variants of this study in ExAC might signify for disease-causing potential.

4.5. Discussion

4.5.1. WES study considerations in uncovering de novo variants

The WES study presented in this chapter was based on the hypothesis that *de novo* point mutations and indels significantly contribute to the incidence of TGA. In previous studies, focused on familial TGA, a number of rare variants were uncovered in laterality genes *ZIC3*, *CFC1* and *NODAL* (Unolt *et al.*, 2013). These studies established TGA as a disease which can be caused by one or a small group of mutated genes. However, the measured recurrence rate in relatives of patients with TGA across studies remains low, with most cases of TGA considered sporadic (Digilio *et al.*, 2001; Unolt *et al.*, 2013). The overwhelming share of sporadic cases could be explained to some extent by the deleteriousness of the disease. TGA is a highly lethal CHD, with 95% of patients dying within the first year of life when the transposition is not surgically corrected (Saremi, 2014). With only recent generations having benefitted from corrective surgery, most mutations predisposing to TGA having arisen in patients in the past are expected to have been eliminated by natural selection. *De novo* variants are thus suggested to play a large role in sporadic TGA. Another hypothesis is that environmental factors play the larger role in the development of TGA. In Section 4.2.1, I mentioned some of the known environmental triggers.

Studying *de novo* variants using WES presents its own series of challenges. As in other WES studies, variants might be missed due to low quality reads and low read depth (Altmann *et al.*, 2012). This is discussed in chapter 1 and again in chapter 3. In this context, the issue extends to whether a variant detected in the proband can be considered *de novo*. Given perfect sequence coverage, a variant can be considered *de novo* if is detected in the proband but not in either parent. In reality however, variant sites may have too few reads in one or both parents, making a genotype assignment difficult. Filtering such cases out, as has been done in the present study, can exclude genuine *de novo* variants. Another possible approach is to rely on other types of filtering to differentiate *de novo* variants from inherited variants. This can be done for example by using a stringent MAF threshold. This approach presents its own problems. Population databases, such as the ESP (NHLBI, 2015) may contain miscalls reported as a low frequency variants. It is also possible that a few of these rare variants will match *de novo* variants. Additionally, using a stringent MAF will not be able to filter out variants that are not *de novo* but have nonetheless occurred as a result of a recent mutation event. In chapter 1, I reviewed the literature around the distribution of variants in the human population which found an abundance of rare variants compared to what would be expected from population genetic theory (Altshuler *et al.*, 2010). This will include many variants not present in either the 1000g or ESP database.

As shown in Table 10, the number of variants detected as *de novo* was in excess of the number of variants expected. True *de novo* variants had to be extracted from this list. In Section 4.2.2, I mentioned the rate of *de novo* variants as being 0.74 in the exome per individual. Provided *de novo* variants are indeed causing TGA, the actual rate of *de novo* variants per exome will be slightly higher due to selection bias. However my results from *de novo* detection are still one order of magnitude higher. This will once more largely be imputable to sequencing error and the incorrect assignment of genotypes by SAMtools and GATK. Mosaicism in parents might also contribute to this effect. Mosaicism refers to the presence of cell lines with different genotypes at a given locus within a single individual due to a postzygotic de novo mutation event (Biesecker and Spinner, 2013). A variant could have therefore been detected as *de novo* in a proband whilst being inherited if it is not present in the blood cells from the parents; or appears with low frequency, in the parent that carries the variant (Biesecker and Spinner, 2013). Conversely, a postzygotic *de novo* mutation event can lead to mosaicism in the proband. The resulting variant could thus cause the manifestation of TGA while not being detected in DNA from blood samples. It is difficult at this time to determine the contribution of somatic mutation to CHD as few studies have investigated this possibility. In one recent study of cardiac tissues and blood samples from 52 unrelated patients with ToF, Huang *et al*. (2013) identified two somatic variants in *GATA6*; a gene previously implicated in CHD, in patients. To my knowledge, no somatic variants have been reported as a trigger for a case of TGA yet.

The rate of variants identified in my study once filters have been applied is at around 0.75 per proband. This is close to the rate of *de novo* variants found in the literature as 0.74. Validation work suggests that this final list of variants does in fact mostly contain genuine *de novo* variants, with only 1 of 15 variants sent for validation not confirmed to be *de novo*; with another variant pending validation. However, the difficulty is in knowing how many *de novo* variants were missed as a result of sequencing and filtering. It is possible that the rate of actual *de novo* variants in the 32 probands is higher. There is also little indication as to which of these variants could be disease-causative beyond functional predictions. Variants were not found in genes already associated with TGA or heterotaxy. Likewise, no two variants were found in the same genes. This last point must be contrasted with the sample size used for this study. A total of 32 proband were studied, with consensus list variants found in 14 probands, some of which appear to be unlikely to be disease-causative. In their study of the contribution of *de novo* variants in severe sporadic CHD, Zaidi *et al*. (2013) estimated that a total of 400 genes, when mutated, could be contributing to severe sporadic CHD. The consequence of such locus heterogeneity for small studies is that overlap in genes with disease-causing *de novo* variants would be unlikely to occur. If their conclusions can be applied to single CHD subtypes; in this case TGA, this could explain why my study only does not include two *de novo* variants in the same gene. In the following section, I describe the *de novo* variants predicted to be deleterious found in 11 probands.

4.5.2. De novo variants by family

The WES study focused on 32 families. The final consensus list contained 24 variants across 14 probands. 16 variants in 11 probands were predicted to be probably deleterious; or equivalent, by at least one functional predictor.

In the proband of family 33, a total of 4 candidates were detected by SAMtools and GATK. The variants are distributed among four genes: *ANGPTL2*, *KCNJ12*, *LMOD3* and *ZNF419*. Only variants in *ANGPTL2* and *KCNJ12* were predicted as probably deleterious. It was not possible to establish whether the variant in *KCNJ12* is a *de novo* variant in the proband. However, using ExAC, I determined that the variant, although rare was unlikely to be a *de novo* given it was identified in 229 other individuals. At the very least, its presence across this many individuals suggests that it is not disease-causing. *ANGPTL2* appears to be the most likely to be pathogenic. According to RefSeq annotation (Pruitt *et al.*, 2014), *ANGPTL2* is part of an ensemble of angiopoietins, proteins expressed almost exclusively in the vascular endothelium. Angiopoientins are responsible for the formation of new blood vessels; or angiogenesis, and thus are part of the vascular endothelial growth factor (VEGF) family. *ANGPTL2* is also abundantly expressed in adult heart and embryonic atrioventricular canal cell tissues (Rebhan *et al.*, 1998; Kim *et al.*, 1999). Possibly underscoring a role in the formation of the great arteries is the mouse homolog gene *Angptl2*'s role in the inflammation leading to abdominal aortic aneurysm (AAA) (Tazume *et al.*, 2012). Tazume and his team were able to demonstrate an abundant expression of the Angptl2 protein in macrophages within vessel walls and to show a reduction of aneurysm size and vessel structure destruction in mice with Angptl2-deficient macrophages. Mice knockouts in *Angptl2* display a reduced distribution of microphages at inflammatory sites (Eppig *et al.*, 2015). Based on the current evidence, the best candidate in this family appears to be the variant in *ANGPTL2*.

In family 70, a candidate was found in gene *CPT2* and predicted as disease-causing by MutationTaster, but benign; or neutral, by others. *CPT2* encodes an enzyme that contributes to the transfer of fatty acid from the cytosol to mitochondria for oxidation (Longo *et al.*, 2006) The enzyme produced by *CPT2* is active within the mitochondrial inner membrane (Longo *et al.*, 2006). To date, genes involved in cardiogenesis have not been found to act through mitochondria (Gelb and Chung, 2014). From this I have concluded that the gene is unlikely to play a role in cardiogenesis.

In family 85, one candidate was found in *ZNF577*, a gene encoding for a zinc finger protein. Pathogenicity predictions are mixed. Few annotations exist for this gene that could be used to either support or rule out a role in cardiogenesis. The distribution of truncating variants in ZNF577, as seen in ExAC, suggests that the abnormal activity of *ZNF577* could be tolerated by the organism.

In family 216, one candidate was found in *RBP5* and predicted as disease-causing by MutationTaster, but benign otherwise. *RBP5* encodes for a retinol binding protein, retinol being the alcoholic form of vitamin A (Folli *et al.*, 2001). According to Folli *et al.* (2001), gene expression in adult is largely specific to the kidney. However, the identification of a *de novo* variant in a gene that encodes a retinol binding proteins remains interesting. TGA has been reported as associated with high maternal intake of retinol supplements (Loffredo *et al.*, 2001). Retinoic acid, the active form of vitamin A, is used to induce TGA in newborn mice via ingestion in the mother (Unolt *et al.*, 2013). Retinoic acid antagonists are also used to induce TGA, suggesting that normal cardiac development depends on a balance in the level of retinoic acid (Cipollone *et al.*, 2006). Additionally, Nash *et al*. (2015) identified a variant in *RBP5* shared between 5 distantly related patients with total anomalous venous return (TAVR) that they suggests contributes to the disease. Briefly, a TAVR occurs when all pulmonary veins are malpositioned during cardiogenesis, leading to the wrong connections being made. While TAVR and TGA are two distinct CHDs, these have been observed together, which could be indicative of some shared etiology (Raff *et al.*, 2002). There are a number of reasons to be cautious about the results presented by Nash *et al*. (2015). Firstly, the variant has a MAF of around 9%. According to the authors, the variant is overrepresented in patients with TAVR, but not in patients with heterotaxy. Secondly, the gene selected for knockdown in zebrafish, *rbp7a*, only shares 51% sequence similarity with RBP5. A little under half of the mutant zebrafish display abnormal right-side looping of the heart which, at the very least, establishes a role for retinol binding proteins in cardiogenesis (Nash *et al.*, 2015). On this basis, a role for the *RBP5 de novo* variant found in this family's proband should not be excluded. It can be hypothesised that a mutated *RBP5*, leading to a malformed protein that cannot bind retinol as effectively as a normal protein, could decrease the level of retinol available to the cell in the developing heart. This would then lead to less retinoic acid, with the mutation having the effect of a retinoic acid antagonist.

In family 222, two candidates were found, in *FAM208B* and *PROK1* respectively. The variant in gene *FAM208B*; for which few resources exists, is predicted to be benign. On the other hand the variant in *PROK1* is predicted as pathogenic by all four pathogenicity predictors used in this study. *PROK1*, which encodes a prokineticin protein, belongs to the VEGF family (LeCouter *et al.*, 2003). This is the second variant identified in a gene belonging to the VEGF family in the consensus list with *ANGPTL2*. As its early name, endocrine-gland-derived vascular endothelial growth factor, suggest, it is specifically expressed in endocrine tissue in the ovary, testis, placenta and adrenal gland (LeCouter *et al.*, 2003). This targeted expression makes *PROK1* an unlikely actor in cardiogenesis and by extension, unlikely to be a cause of TGA when mutated.

In family 245, the candidates were found in *DCDC1*, *GREB1* and *RAD52*. The variant in *DCDC1* is largely predicted to be benign with some prediction of deleteriousness for the other two genes. *RAD52* encodes a protein that plays an integral role in double-strand break repair and DNA recombination (Park *et al.*, 1996). This specific role suggests that the protein probably does not play a role in cardiogenesis. *GREB1* is characterised by its high expression in estrogen-receptor-positive breast tumors (Ghosh *et al.*, 2000). It is also expressed in prostate cancer tissue (Rae *et al.*, 2006). Relevant to this study is its expression in atrioventricular canal cells in embryonic tissue, a characteristic it shares with *ANGPTL2* (Rebhan *et al.*, 1998). However, the strongest evidence for a potential role for *GREB1* in causing TGA is the previous identification of a CNV in a patient with malposition of the great arteries (MGA) which overlaps this gene (Fakhro *et al.*, 2011). Fakhro *et al*. (2011) tested the expression of GREB1 in *Xenopus Tropicalis* using in situ hybridization and found expression patterns that suggested that the gene could indeed play a role in left-right patterning; and cardiogenesis in general (Rebhan *et al.*, 1998; Kim *et al.*, 1999). The subsequent gene knockdown conducted by Fakhro *et al*. (2011) did not lead to an abnormal left-right looping phenotype. Despite this, *GREB1* remains a strong candidate in this present study. An interesting finding using ExAC was the identification of 11 other unrelated individuals with the same variant. It is possible that these individuals were part of one of the disease-specific population studies incorporated by ExAC, although this cannot be confirmed at this time (The Broad Institute, 2015a). None of the studies focus on CHD, but it is possible that some patients with multiple defects including CHD would have been incorporated to a study on the basis of the extracardiac defects.

In family 280, two candidates were found, in *PHLPP2* and *ZNF227* respectively. Both candidates have a high disease-causing potential, with the variant in *ZNF227* being a truncating variant. *PHLPP2* encodes a protein which is known to mediate dephosphorylation of several genes (Brognard *et al.*, 2007). Of particular interest is the action of *PHLPP2* on *AKT1*, a gene embedded in the VEGF signalling pathway (Brognard *et al.*, 2007). *PHLPP2* acts to suppress *AKT1*'s activity (Brognard *et al.*, 2007). *PHLPP2* is thus a third gene in the final list of variants found to be involved with the VEGF pathway. As with the *ZNF577* variant, there are few annotations about *ZNF227*. The *de novo* variant in this case is a nonsense variant. However, the distribution of truncating variants seen with ExAC suggests that truncating variants in this gene might not cause a deleterious phenotype.

In family 311, a candidate nonsense variant was found in *GDPGP1*. *GDPGP1* regulates GDP-D-glucose levels in cells (Adler *et al.*, 2011). Few annotations exist about this gene. The variant in *GDPGP1* was validated as an actual *de novo* variant, but has also been identified in an individual reported in ExAC. Additional information on the patient identified in ExAC could either support or rule out this variant as a cause for TGA. Given the specific nature of the protein function, a role in cardiogenesis is unlikely.

In family 312, two candidate variants were found, in *PDE4D* and *ENTPD2* respectively. Both appear to be strong candidates based on functional prediction, the latter being a frameshift insertion. PDE4D acts on cyclic AMP (cAMP) (Houslay *et al.*, 2007). The cAMP molecule acts as a regulator for different cell processes, with specialised functions in different cell types (Houslay *et al.*, 2007). *PDE4D,* a member of the phosphodiesterase family, in turn, regulates the action of cAMP (Houslay *et al.*, 2007). Different isoforms are expressed in different tissues, with isoform 7 being most expressed in heart and skeletal muscle. *PDE4D* appears to be associated with heart failure, but not severe cardiac defects arising during cardiogenesis (Houslay *et al.*, 2007). *ENTPD2* encodes for an enzyme in the cell membrane of the ectonucleoside triphosphate diphosphohydrolase family (Chadwick and Frischauf, 1997). The gene is expressed in atrioventricular node cells in the embryonic heart and in the adult heart (Rebhan *et al.*, 1998). The presence of a frameshift insertion makes *ENTPD2* an interesting candidate, but there seems to be little literature hinting at a role for *ENTPD2* in cardiogenesis.

In family 328, two variants were identified, one in *COLL1A2* and the other in *GPR17*. Both variants are largely predicted as benign with the exception being an assignment of 'deleterious' by LRT for the variant in *COL11A2*. *COL11A2* encodes an alpha chain for type XI collagen (Lui *et al.*, 1996). In turn, type XI collagen is important for skeletal integrity (Lui *et al.*, 1996). Diseases associated with mutations in *COL11A2*, such as Stickler syndrome, largely involve skeletal defects and hearing loss (Rebhan *et al.*, 1998). Around half of patients with Stickler syndrome also have a regurgitating mitral valve (Liberfarb *et al.*, 1986). There does not appear to be any described cases of Stickler syndrome with TGA. As established earlier, TGA is seldom seen with extracardiac defects (Unolt *et al.*, 2013).

In family 503, two candidates were found in *CREB3L3* and *SNX13* respectively. The variant in *CREB3L3* is predicted to be benign. The variant in *SNX13* is predicted to be deleterious by Polyphen2 and LRT. *SNX13* belongs to the sorting nexin family of proteins and is involved in intracellular trafficking (Rebhan *et al.*, 1998). Interestingly, knockout of the homolog gene in mice can lead to abnormal blood vessel morphology and is lethal to the mouse embryo (Zheng *et al.*, 2006). Another member of the sorting nexin family, *SNX10*, has been found to act in a regulatory pathway for ciliogenesis; the development of cell cilia (Chen *et al.*, 2012). Although still highly speculative, this presents the possibility of a role for mutated sorting nexins in ciliopathy; diseases of dysfunctional cilia. The implication for TGA being that disorders relating to abnormal cilia can result in heterotaxy, a feature that accompanies some cases of TGA (Ware *et al.*, 2011).

In family 540, the candidate is in gene *ZFHX3* and is predicted as potentially deleterious by LRT and MutationTaster. *ZFHX3* encodes a transcription factor with several homeodomains and zinc finger motifs (Rebhan *et al.*, 1998). The gene has been linked to atrial fibrillation, but not any form congenital cardiac defect (Benjamin *et al.*, 2009).

The variants in these families do not occur in heterotaxy genes or genes associated with TGA. There does not appear to be any unifying theme to categorise these variants. Several are involved in the VEGF pathway, but this association does not reach significance. Furthermore, one of the genes in this pathway, *PROK1,* is specific to tissues that did not include the heart. A few variants occur in genes that could play a role in TGA based on evidence from previous experiment. This is the case for example of variants in *RBP5* and *GREB1*. A role for *SNX13* also seems possible. No two candidates were found in the same gene; however the cohort was relatively small in size. A larger cohort of sporadic TGA cases might help further elucidate the present cases. Variants found in these same genes in other WES studies of sporadic TGA could lend credibility to some of the candidates identified in this study. This extends to the study of *de novo* CNVs in TGA patients which might also reveal CNVs affecting the same genes identified in this study. Finally, it cannot be ruled out that the scarcity of candidates could be an indication that TGA is largely caused by environmental factors.

4.6. Conclusion

The WES analysis of 32 families; 31 parent-offspring trios and a single quartet, has revealed a number of *de novo* variants distributed across 14 of the probands affected by TGA. The rate of *de novo* per proband is close to what would be expected with 0.75 *de novo* variants detected per proband. Among 24 *de novo* variants detected to date, 16 are predicted to be probably deleterious; or equivalent, by at least one pathogenicity predictor and 14 have been validated. None of the variants detected occur in genes previously associated with TGA or heterotaxy. The genes were also not part of a list of 276 genes highly expressed in heart compiled by Zaidi *et al*. (2013). There was no unifying theme associating a subset of the genes found with disease. Despite all this, the existing literature suggests that the variants that occur in several genes could be disease-causative. These include the variants in *GREB1*, *RBP5* and *SNX13*. Other variants identified in this study could have an impact on the development of TGA through some as of yet unknown mechanism. The genes in which these variants were found can be compared to those that appear in future studies. The re-occurrence of any of these genes in future WES studies of TGA patients would be further evidence that these genes, when mutated, are involved in ToF. If TGA can be caused by *de novo* mutations in many distinct genes, large trio studies will be required to better determine which genes, when mutated cause ToF.

Chapter 5. Whole-exome sequencing study of 824 patients with Tetralogy of Fallot

5.1.1. **Summary**

Tetralogy of Fallot (ToF) is a complex form of congenital heart disease (CHD) combining four heart morphological defects: a large ventricular septal defect, a displacement of the aorta over the septal defect, a narrowing of the pulmonary valve and a thickening of the right ventricular wall. The most common form of cyanotic CHD, its genetic etiology remains nonetheless largely unresolved. In this chapter, I present a large-scale WES study involving 824 unrelated ToF patients with sequencing performed at McGill University and the Génome Quebec Innovation Centre (MUGQIC). This study focuses on identifying genes that potentially play a role in the development of ToF using various approaches. In a first stage, I categorised rare variants according to their potential to disrupt normal protein function. I compared the genes in which these rare variants occur to sets of genes previously associated with CHD in the literature. In a second stage, I identified clustering of rare variants across the length of gene coding sequences (CDS) and exons in the 824 ToF cases. I hypothesised that rare variants influencing the disease trait would cluster in ToF patients. The genes in which clustering occurred were also compared to sets of genes implicated in CHD. In the final stage of my study, I conducted a pathway analysis using STRING v.10 (Jensen *et al.*, 2009) and Enrichnet (Glaab *et al.*, 2012). With STRING, I checked the list of genes in each cluster category for an enrichment of protein-protein interactions. I then used Enrichnet to determine if the genes fell within any specific pathways or processes. Using interaction data from STRING, I looked for genes in each cluster category that might directly interact with known ToF genes. To test the assumptions made at every step of the study, I used a second set of 490 cases with various neurodevelopmental disorders, obtained from the UK10K project, as comparators. This study reveals that rare single nucleotide variants (SNVs), truncating or otherwise predicted to be deleterious, in known ToF genes are found in 4.4% of patients with ToF. This suggests that genes previously identified as influencing ToF through single highly-deleterious variants only account for a small fraction of ToF cases. The study of SNVs clustering in ToF patient revealed the novel candidate *FLT4* and established *NOTCH1* as contributors to ToF. Overall, results of the study suggest that, despite the identification of both known and potentially novel genes influencing the risk of ToF, single highly-penetrant variants do not play a major role in ToF, implying a more complex genetic etiology is at work in many patients presenting ToF.

5.2. Introduction

5.2.1. Tetralogy of Fallot

As described in Chapter 1, ToF is one of the most common forms of CHDs accounting for 7-10% of all CHD pathologies (Bailliard and Anderson, 2009). It is the most common form of cyanotic CHD which is characterised by an intermittent or permanent bluish appearance as a result of low blood oxygenation (Apitz *et al.*, 2009). The principle abnormality is an antero-cephalad deviation of the ouflow tract septum, resulting in four linked phenotypic features: a ventricular septal defect (VSD), a displacement of the aorta over the VSD, a narrowing of the pulmonary valve; described as pulmonary stenosis (PS), and a progressive thickening of right ventricular wall, or right ventricular hypertrophy (RVH) (Bailliard and Anderson, 2009; Nelson *et al.*, 2014). The VSD occurs in the anterior part of the muscular ventricular septum and is typically quite large (Lev and Eckner, 1964). As a result of the septal defect and of an aorta overriding both ventricles, blood circulation is perturbed, with oxygen poor blood that has not traversed the pulmonary circulation being ejected to the systemic circulation, a phenomenon described as shunting (Lev and Eckner, 1964). The PS, which can occur at different levels along the pulmonary outflow tract, also greatly perturbs blood circulation, restricting blood flow through the lungs (Anderson and Weinberg, 2005). Stenosis is occasionally associated with the incomplete development of the pulmonary valve, described as hypoplasia, with either a fusion of leaflets; pulmonary atresia (PA), or a complete absence of the pulmonary valve (Bailliard and Anderson, 2009). The absence of the valve leads to blood flowing back into the right ventricle. The right ventricular hypertrophy is considered a secondary phenotype, developing progressively as the right ventricular muscle works to compensate for the PS (Nelson *et al.*, 2014) . In addition to the four principal morphological defects that characterise ToF, there exist a number of additional minor cardiac defects. Bailliard and Anderson (2009) report that a quarter of ToF patients have their aortic arch on the right. Half of patients with PA have a persistent patent ductus arteriosus which provides blood with access to the pulmonary artery via the aorta (Bailliard and Anderson, 2009). In some cases, the large VSD characteristic of ToF is accompanied by additional smaller septal defects, both of the ventricles and atria. The tricuspid valve is occasionally found to be displaced over both ventricles in ToF patients (Bailliard and Anderson, 2009).

Arthur Louis Étienne Fallot, described ToF as the abnormality underlying the majority of cases of “la maladie bleue” (Anderson and Weinberg, 2005). Cyanosis is the main physical sign associated with ToF, although it is absent from some mild cases (Lev and Eckner, 1964). The degree of cyanosis depends on the level of severity of the underlying defects, particularly the extent of the PS. In a clinical study of infants with ToF, Shinebourne *et al*. (1975) describe cases of persistent cyanosis. The authors point to pronounced infundibular stenosis, which involves a diminished pulmonary valve ring, as the root cause of the severity of cyanosis in these patients. Shinebourne *et al*. (1975) also describe patients with intermittent cyanosis, brought on by crying, which gradually increases over time. Other outcomes of ToF include loss of consciousness, heart murmurs, shortness of breath and eventually, heart failure (Shinebourne *et al.*, 1975). Without surgery, 80% of affected individuals die before the age of 10 (Starr, 2010). However, thanks to advances in modern surgery, patients who undergo complete surgical repair in early life show an 85% survival rate over the first 30 years of life (Bailliard and Anderson, 2009).

The disease etiology of ToF has yet to be fully uncovered. ToF has been reported as a component of several syndromes; diseases characterized by several defects in distinct tissues and organs. These include patients with chromosomal aneuploidy, such as trisomies of chromosomes 13, 18 and 21, as well as large copy number variants (CNVs) inducing disorders such as 22q11.2 deletion syndrome (Nelson *et al.*, 2014). A 22q11.2 deletion has been estimated to be present in 15% of ToF patients, a proportion that rises to 40% when considering patients with PA (Nelson *et al.*, 2014). Although it is important to note that 22q11.2 deletion syndrome refers to a range of phenotypes that do not all include ToF or CHDs in general. When 22q11.2 deletion syndrome does cause ToF, the defect is often accompanied by several extra-cardiac defects.

A few patients with isolated ToF; unaccompanied by extra-cardiac defects, have been described as members of apparently Mendelian families. For example, a missense variant in *JAG1* was shown to follow an autosomal dominant pattern in a family with multiple CHD phenotypes, including ToF (Eldadah *et al.*, 2001). No family member presented Alagille syndrome, typically caused by mutation in *JAG1* (Eldadah et al., 2001). However, the majority of ToF cases studied are sporadic. Excluding ToF patients with 22q11.2 microdeletions, Burn *et al.* (1998) found a recurrence risk of CHD in siblings of 2.2% and of 3.1% in offspring hinting at a strong, but complex, genetic component underlying isolated non-Mendelian ToF (Burn *et al.*, 1998). A genome-wide association study (GWAS) carried out on 839 ToF patients; some of which are included in the study in this chapter, and 5159 controls, conducted by Cordell *et al.*, (2013b)**,** suggests that common variants influence the risk of ToF. This includes single nucleotide polymorphisms (SNPs) in regions 12q24 and 13q32, the latter SNPs coinciding with the *GPC5* locus. The identification of common variants influencing the risk of ToF in region 12q24 is consistent with a candidate gene study, carried out on some of the same patients, which revealed an association between a common variant in *PTPN11* and the disease (Goodship *et al.*, 2012). However, as Cordell *et al.* (2013b) point out, ToF is associated with a high and early mortality rate and therefore severely reduced reproductive fitness. Alleles conferring additional risk of an individual developing ToF would be expected to be constrained by natural selection to allele frequencies lower than the ~5% which typically yield GWAS signals. We therefore expect rare and *de novo* genetic variants to play a substantial role in the development of ToF. Greenway *et al.* (2009) identified 11 rare CNVs in ToF patients that were either absent or extremely rare (>0.001) in a set of controls. Greenway *et al.* (2009) highlight their discovery of CNVs, predominantly duplications, occurring in region 1q21.1, in 5 of 114 ToF patients and none of their controls. The association between duplications in region 1q21.1 and ToF was later confirmed by the work of Soemedi *et al.* (2012a) and Silversides *et al.* (Silversides *et al.*, 2012). Small duplications coinciding with the *GJA5* locus, suggests this gene plays a role in the development of ToF and other forms of CHD (Soemedi *et al.*, 2012a). Results from case-control studies of CNVs in ToF suggest that large CNVs are significant contributors to the disease burden in ToF (Silversides *et al.*, 2012; Soemedi *et al.*, 2012a; Soemedi *et al.*, 2012b). This is true whether or not one chooses to include patients with a 22q11.2 deletion.

To date, few genes have been conclusively associated with ToF. Studies attempting to uncover genes that contribute to ToF through indels and SNVs have largely focused on sequencing genes thought to be good candidates. This typically includes genes that coincide with known CNVs in ToF such as *TBX1* which resides in the 22q11.2 region, and genes previously associated with CHDs (Griffin *et al.*, 2010). Table 18 provides a summary of genes associated with ToF to date through gene sequencing studies.

Gene	Patients with rare variant	Size of control set	Reference
NKX2-5	9/201*	50	(McElhinney *et al.*, 2003)
ZFPM2	2/47	120	(Pizzuti *et al.*, 2003)
GATA4	1/201*	159†	(Tomita-Mitchell *et al.*, 2007)
CFC1	8/121*	125	(Roessler *et al.*, 2008)
FOXH1	5/121*		
TDGF1	2/121*		
GDF1	3/121*	125	(Roessler *et al.*, 2009)
NODAL	12/121*		
JAG1	2/94*	100	(Bauer *et al.*, 2010)
TBX1	3/93	500	(Griffin *et al.*, 2010)
GJA5	2/178	784	(Guida *et al.*, 2013)
GATA6	2/52	200	(Huang *et al.*, 2013)
GATA5	2/35*	200	(Jiang *et al.*, 2013)
BVES	4/114	400	(Wu *et al.*, 2013)
FOXA2	4/93	500	(Topf *et al.*, 2014)
FOXC1	4/93		
FOXC2	2/93		
HAND2	1/93		
NKX2-6	1/43*	200	(Zhao *et al.*, 2014)
TBX5	2/94	200	(Baban *et al.*, 2014)

Table 18. Gene sequencing studies that report a causal link between a gene and ToF. *Study includes case patients with different CHDs, the number reported is the number of case patients with ToF. †105 additional controls were tested for two specific variants in exon 6.

Given the heterogeneity of loci causally linked with isolated ToF, extracting genetics insights from WES studies rather than through individual candidate gene sequencing calls for a large set of cases and controls.

5.2.2. The study of disease with locus heterogeneity using WES

Many rare diseases are caused by single highly-penetrant rare variants occurring within a specific locus. While ToF is a common cardiac defect, it can be hypothesised that many cases of the disease are also result from single highly-penetrant rare variants. The distinction with rare disease is in the number of different loci at which the disease-causing variant may arise, a phenomenon referred to as locus heterogeneity (Bamshad *et al.*, 2011). 20 genes have already been causally linked to ToF, as shown in Table 18. To identify the gene regions in which rare variants causing a particular disease occur, a large set of unrelated patients that share the disease is needed. Given a large set of cases, variants contributing to a disease trait will form detectable clusters within specific regions of a gene or genes which, when mutated, lead to the disease. Given reliable sequencing data, clusters can be uncovered using the W_d statistic (Lange, 1997). There will also be a higher frequency of case patients with variation in genes contributing to the disease compared to a random sample of the population. To capture this frequency increase requires a reference set that can represent this random sample of the population.

The development of NGS techniques, particularly WES, provides researchers with new opportunities in elucidating the genetic etiology of diseases such as ToF, although it also presents its own challenges (Bansal *et al.*, 2010). Disease-causing rare variants have to be extracted from the background of human variation and sequencing error. For any NGS study, filtering criteria to extract genuine rare variants of functional significance from the rest of variation are essential (Altmann *et al.*, 2012). Filtering criteria need to be consistent across cases and controls in order to avoid introducing bias that may confound the results of a study. Ideally, case and control samples need to be sequenced together. This avoids the kind of bias that arises from using different sequencing and capture technologies and protocols for cases and controls (Derkach *et al.*, 2014). It also avoids sequencing bias that may arise from the same sequencing platform being calibrated differently (Derkach *et al.*, 2014). Differences in read depth between cases and controls can lead to differences in MAF estimation and, as a consequence, to spurious associations while true associations are missed (Derkach *et al.*, 2014). Derkach *et al.* (2014) provide a method, the Robust Variance Score (RVS), to address these biases. With this method, Derkach *et al.* (2014) aim to make publicly available genome-wide data suitable as controls for large NGS case-control studies. Using large public control resources across studies could have a transformative effect, similar to what was seen in GWAS as first illustrated by The Wellcome Trust Case Control Consortium (WTCCC) study which compared 14,000 cases from 7 common diseases to 3,000 controls (Burton *et al.*, 2007). However, the efficacy of this method remains to be confirmed through independent studies.

In NGS case-control studies, testing whether a single rare variant is associated with a disease trait requires a number of sequenced samples in excess of what can often be practically achieved. For example, Lee *et al.* (2014b) estimate that 460 individuals would have to be sequenced in order to ensure a 99% probability of sampling an allele with a minor allele frequency (MAF) of 0.5%. Single variant testing becomes particularly problematic if a number of the variants associated with a disease trait are expected to be private, in other words, only in a single individual. Another thing to remain mindful of is that there are large numbers of rare variants in the human genome, as confirmed by Tennessen *et al*. (2012), which will not be associated to disease. The study I present in this chapter is based on the assumption that a major cause of isolated sporadic ToF are single highly-penetrant rare variants. However, for many diseases, including some subtypes of CHD, a more complex etiology has been proposed over the years, leading to the methods that I will now present (Gelb and Chung, 2014). Given a CHD with a more complex genetic etiology of ToF, one of the following methods could form the basis of a future WES study design.

The alternative to single variant testing mentioned above is to perform a multiple variants test. For this kind of test, distinct variants are aggregated according to some feature, such as a shared gene region or functional relevance (Bansal *et al.*, 2010). This process is often referred to as collapsing (Dering *et al.*, 2011). Collapsing involves a summarisation of the presence of rare variants within some unit of interest in both case and control samples (Bansal *et al.*, 2010). All collapsing methods can be refined by selecting which type of variant to collapse. For example, collapsing can focus on rare truncating variants or variants predicted to be deleterious in genes or groups of genes (Dering *et al.*, 2011). Lee *et al*. (2014b) describes five broad categories of multi variant tests: burden and adaptive burden tests, variance-component tests, omnibus tests and the exponential combination test. The first four sets of tested have been implemented in a number of methods working on different assumptions about patient sequence data (Lee *et al.*, 2014b).

Burden tests were the first to be devised, with Mogenthalier and Thilly's (2007) 'Cohort Allelic Sums Test' (CAST). CAST represent collapsing in its simplest form: a binary value is assigned to each gene region in an individual depending on the presence or otherwise of rare variants in that region (Lee *et al.*, 2014b). The results of CAST over cases and controls can be summarised as a score using Fisher's exact test (Moutsianas and Morris, 2014). The assumption made here is that the mere presence of rare variants in a region increases disease risk (Lee *et al.*, 2014b). To test a dominant genetic model; where each rare variant in some region contributes to the disease risk, the binary value becomes a count of the number variants in the region, a model underlying the MZ test by Morris and Zeggini (2010). The disadvantage of this approach is that any rare variant that has no effect on the disease trait will interfere with the signal of any real association. The aforementioned methods require a MAF threshold above which variants are excluded from the test (Lee *et al.*, 2014b). The presence of variants that are too common and therefore likely to be in cases and controls may interfere with the signal emanating from large effect rare variants associated with the disease. Setting a MAF threshold can lead to the exclusion of causal variants if too low or the inclusion of many non-causal variants if too high (Moutsianas and Morris, 2014). Both will reduce the test's power. Putative differences in effect can be partially captured through a variant's rareness. With their 'Weighted Sum Test' (WST), Madsen and Browning (2009) replace the MAF threshold with a weighting system which varies as a function of a variant's known MAF. Rarer alleles will be given a higher weight. The attribution of weights to variants is further explored by the 'Combined and Multivariate and Collapsing' (CMC) method which attempts to capture the effect of both rare and common variants on a disease trait, assigning weight in intervals (Li and Leal, 2008).

Burden test have largely been superseded by adaptive burden tests, a number of methods that account for the possibility that specific variants within a collapsed region can have trait-decreasing effects or no effect at all (Lee *et al.*, 2014b). Given that the actual direction of the effect of each variant is not known, estimates are made through a marginal model. In these tests, such as Han and Pan's (2010) aSum test, variants that are estimated to have a trait-decreasing effect are given a negative weighting (Moutsianas and Morris, 2014). One of the pitfalls of these methods however is that, in order to determine the statistical significance of a possible association, the test has to be carried out on every possible marginal model arising from a different distribution of the trait amongst individuals tested (Moutsianas and Morris, 2014). In other words, the model needs to account for all the possible phenotype distributions across a set of individuals. This makes scaling up the tests to the whole-exome or whole-genome impractical, if not impossible (Moutsianas and Morris, 2014). This issue is addressed by variance-component tests, also referred to as dispersion tests (Lee *et al.*, 2014b; Moutsianas and Morris, 2014). Unlike burden tests, variance-component tests were designed to account for trait-decreasing and neutral effects (Lee *et al.*, 2014b). Variance-component tests analyse the variance in genetic effects between variants in a group (Lee *et al.*, 2014b). A test statistic is produced for each variant and it is the variance in these statistics that is analysed (Lee *et al.*, 2014b). In the `Sequence Kernel Association Test' (SKAT), the distribution of effects is compared with what would be expected under the null hypothesis of no association with a disease trait (Wu *et al.*, 2011). As with burden testing, different methods built around variance-component exists to address different assumed models of inheritance. To study a single trait with no covariates, the C-alpha test can be used (Neale *et al.*, 2011). While variance-component tests provide their best results when a small fraction of variants influence the trait and are better equipped than burden tests to deal with trait-decreasing variants, burden tests still offer the best results when most variants within a specific region are trait-increasing (Lee *et al.*, 2014b). Omnibus tests combine burden and variance-component tests, in an attempt to capture the advantages provided by both types of test (Lee *et al.*, 2014b). However, this makes omnibus tests computationally intensive.

In this chapter I present a WES case study involving 867 patients with ToF. Selecting 824 cases for which variant calls presented high specificity, I first looked for rare variants called in genes previously implicated in ToF or CHD. This was followed by an analysis of the clustering of rare variants across the coding sequence (CDS) of genes and exons. Finally, I subjected genes containing rare variants to pathway analysis. At every stage in the case study, results were compared to those obtained in 490 patients with various neurodevelopmental disorders. This study reveals that rare truncating variants in known ToF genes; previously uncovered through gene sequencing studies, are found in 0.8% of patients, and rare variants predicted to be deleterious (excluding truncating variants) in 3.6% of patients. Results from the study of clusters indicate that *FLT4* and *NOTCH1* are overrepresented in ToF patients. Furthermore, evidence suggests that *FLT4* directly interacts with other genes that have been implicated in CHD. At the time of writing, this dataset represented the largest exome-sequencing study of a homogeneous CHD phenotype yet reported.

5.3. Material and Methods

The study initially included a total of 867 patients living in Northern Europe and presenting isolated ToF. DNA extracted from blood and saliva samples were provided by several laboratories working with Newcastle’s Institute of Genetic Medicine (IGM) as shown in Table 19. WES was performed at the McGill University and the Génome Quebec Innovation Centre (MUGQIC) using the Illumina HiSeq 2000 (Illumina, 2014a). The exome capture was performed using Agilent SureSelectXT Human All Exon 50Mb kit (Agilent Technologies, 2015).

Location	Number of Samples
Bristol Royal Hospital for Children, UK	70
CONCOR, Netherlands	63
Leeds General Infirmary, UK	147
Centre for Human Genetics, Leuven University, Belgium	83
Royal Liverpool Children’s Hospital NHS trust, UK	119
Newcastle Royal Victoria Infirmary, UK	127
Institute of Genetics, Nottingham University, UK	72
Department of Cardiovascular Medicine, Oxford University, UK	142
The Children’s Hospital at Westmead, Australia	44

Table 19. Locations from which blood and saliva samples were obtained from 867 ToF patients.

This large series of ToF cases was initially supposed to be accompanied by a sequencing matched reference set. However, we were not able to obtain this sequencing matched set. I therefore acquired sequencing data from the UK10K project European Genome-Phenome Archive (EGA) for use as a comparator sample. Access to EGA was limited to a few studies. UK10K healthy controls were only available as low-coverage whole-genome sequencing (WGS) data. A number of WES disease cohorts were available for use as a comparator sample. These cohorts involved patients from the UK and Ireland, presenting various kinds of neurodevelopmental diseases. The four WES studies I selected include patients with autism spectrum disorder (AD), schizophrenia and mental retardation. These are summarised in Table 20. I selected cases for which the sequence data matched MUGQIC’s sequencing strategy; sequencing and exome capture using Illumina HiSeq 2000 and Agilent SureSelectXT Human All Exon 50Mb kit. This further restricted the number of cases available to me. In total, I obtained 500 neurodevelopmental (NDD) cases to be used as controls in a prospective case-control study. Once I determined that a case-control study would not produce meaningful results due to systematic differences between the ToF and UK10K NDD data, I repurposed the NDD cases as a second set of samples against which to test the assumptions made at each stage of my WES case study.

Study name	Pathologies studied	Number of cases included in my study
UK10K_NEURO_MUIR	psychoses and mental retardation	48
UK10K_NEURO_ABERDEEN	schizophrenia	266
UK10K_NEURO_ASD_GALLAGHER	autism spectrum disorder	72
UK10K_NEURO_EDINBURGH	schizophrenia	114

Table 20. Studies from which NDD cases were acquired. The study UK10K_NEURO_MUIR contains sequence data from the Illumina GAIIx which was consequently excluded.

The work this study involves required large amounts of computing power and storage. I therefore carried out my analysis on two different computer clusters: Lampredi2; hosted at Newcastle University, and Hydra; at Manchester University. The clusters are described in detail in Chapter 2 in Section 2.3.1. I accessed the latter remotely from Newcastle. ToF and NDD sequence data files were provided in the BAM file format and the reads were extracted to produce FASTQ files in order for me to perform my own alignment. It was important for sequence alignment and variant calling to be consistent across all sets. The same process is described in Chapter 4. I performed alignments using BWA (Li and Durbin, 2009), removed duplicates using Picard (The Broad Institute, 2015c), and called variants using SAMtools (Li *et al.*, 2009a), filtering out calls falling outside of the exome capture's target region. In addition to ToF and NDD cases, I downloaded sequence data; in BAM file format, from 875 unrelated individuals present in the 1000genomes project (1000g) (Altshuler *et al.*, 2012)**.** I used the same procedure to get my own sequencing alignment. However, variant calls were restricted to a subset of loci of interest first identified in ToF and NDD cases and described below.

For the purposes of the study, I estimated the sensitivity and specificity of each set of variant calls using population data, a technique described by Houniet *et al*. (2015). In this context, sensitivity corresponds to the rate with which the caller identifies variants at sites where variants are present. Specificity corresponds to the rate with which the caller does not identify variants at sites where variants are absent. A large number of erroneous variant calls can quickly become unmanageable in a large-scale study, masking the signal from genuine variants. I have therefore excluded from the study any cases that presented a specificity of ≤99%. In other words, any case that was likely to include many false positives. This means that the final case study effectively includes 824 ToF cases from 867. Similarly, I retained 490 NDD cases out of 500 for this study. Every stage of the WES study involves this set of cases.

Given that ToF cases and NDD cases were sequenced independently, I set out to determine whether the two sets of cases were comparable and therefore whether NDD could be used as controls. To provide an ideal comparison, presence of a disease trait should be the only characteristic that differentiates a disease case set from a comparator set. Differences in sequencing or set composition can be the source of biases between sets, making an objective case-control comparison difficult. In order to determine this, for every gene I counted the number of cases with at least one high-confidence (Q30 in SAMtools) rare synonymous variant, testing for the hypothesis that no significant difference exists in counts between ToF cases and NDD cases. The comparison was done using Fisher's exact test, producing a set of *p*-values that was then compare to a standard chi-square distribution. To determine whether sequencing bias might disappear with variants called with higher confidence, I repeated the same operation using a higher threshold for confidence (Q60).

For this WES study, I only retained variants called with high-confidence (Q30). Using ANNOVAR (Wang *et al.*, 2010), I annotated each set of variant calls, providing me with the information needed to apply different filtering criteria. I used it to extract rare variants. Variants present in the 1000g and the Exome Variant Server (EVS) with a MAF≥1% were excluded from the study as too common to be of interest (MacArthur *et al.*, 2014). I also excluded variants with a frequency of ≥1% in 418 in-house exomes from other studies. I excluded all variants shared between ToF and NDD cases and any variant appearing with a frequency of ≥1% in either set. In addition to removing common variants, these steps are also expected to remove a number of overrepresented sequencing artefacts. To further remove likely artefacts, I excluded any variant falling within a segmental duplication. Segmental duplications were detected using GenomicSuperDups which is part of ANNOVAR annotation (Wang *et al.*, 2010). As a result of the difficulties inherent in indel calling, indel calls produced by SAMtools were much less reliable than those produced for SNVs. Indels were therefore removed from list of variants. These different filtering steps led to a list of variant sites that I used in every subsequent stage of the study.

Using SAMtools, I called the aforementioned variant sites in 875 cases from the 1000g, thus providing a further list against which to filter results from ToF and NDD cases. Variants present in ToF or NDD cases that were also present in the 1000g at least once were filtered as part of a more stringent filtering criterion. This left only variants unique to either set of cases. Additionally, this step removed artefacts. The list of variants from the 1000g available through ANNOVAR will have had some known artefacts removed from it. The variant sites I called from the 1000g will still carry these artefacts and can thus be used to remove the artefacts that may exist in my list of variants. It is worth pointing out that this filter is particularly stringent and will remove variants that would be estimated to have a MAF~0.05% but could be in fact much rarer due to the limited amount of alleles used to estimate the MAF. In Section 5.4, I annotated the results I obtained following my main filtering pipeline to account for the results of obtained after stringent filtering.

At every stage of the WES study, I divided rare variants into four categories: truncating, predicted deleterious, non-synonymous and synonymous. Here I define variants as truncating if they are predicted to shorten the coding segment of a gene (Rivas *et al.*, 2015). This category includes nonsense variants and splice site variants. The following category; predicted deleterious, includes variants predicted as deleterious by a consensus of pathogenicity predictors, excluding truncating variants. The predictors used are MutationTaster, PolyPhen2 and LRT, as in Chapter 4. In order for a variant to be considered predicted deleterious, each pathogenicity predictor has to predict the variant to be at least possibly deleterious. The non-synonymous category includes all missense variants with the exception of those in the previous category. The synonymous category includes all variants that are synonymous and therefore does not overlap with any of the three other categories. To separate the effects of each category, it was crucial that none of the categories overlap. This was particularly crucial for the truncating and predicted deleterious categories for which deleteriousness was determined through dissimilar approaches.

The first step of this WES ToF study was to establish the presence of rare truncating variants and rare variants predicted deleterious in genes previously implicated in ToF and CHD. Mutated genes that influence the risk of ToF will be more common in ToF patients than in patients with some unrelated disease, a fact that is expected to be reflected in the number of cases with one or more rare variants in known ToF genes. I therefore compared genes harbouring rare variants in 824 ToF cases to two sets of genes implicated in ToF as summarised in Table 21.

Gene set	Disease	Genes in set	Source
Known ToF	ToF (isolated)	20	Gene sequencing studies (see Table 18)
IVA	ToF (isolated)	42	Ingenuity Variant Analysis software™
ToF syndrome	ToF (syndromic)	80	Winter-Baraitser Dysmorphology Database
HRC	CHD	312	HeartRepair consortium
Zaidi	CHD (severe)	276	Study by Zaidi *et al.* (2013)
CHD syndrome	CHD (syndromic)	674	Winter-Baraitser Dysmorphology Database
Cilia	CHD (cilia)	61	Study by Li *et al*. (2015)

Table 21. Summary of gene sets used in the study of variants in genes previously implicated in ToF or CHD.

The first set contained genes that were causally linked to isolated ToF through gene sequencing studies; presented in Table 18, while the second set contained genes identified through Ingenuity Variant Analysis software™ (QIAGEN, 2015) as affecting isolated ToF. One gene in the set identified through Ingenuity Variant Analysis software™, *TGFB1*, was annotated as having a "protective effect" and was therefore excluded. Additionally, I looked at variants in genes associated with syndromes that involved ToF, catalogued by the Winter-Baraitser Dysmorphology Database (London Medical Databases, 2014). I also compared the genes in my study to two sets of genes implicated in various types of CHD, also summarised in Table 21. The first set originated from the HeartRepair Consortium (HRC) (EU FP7 Consortium). The HRC is a European research group investigating cardiac development through various methods. The different consortium members have contributed to a set of candidates, each using their own criteria of selection, but with the unifying theme that the genes were somehow involved in CHD. The second set was compiled for a study on the contribution of *de novo* variants to severe CHD, including ToF (Zaidi *et al.*, 2013). Zaidi *et al.* (2013) selected genes associated to CHD through human and model system studies. As before I compared genes harbouring variants in my study to genes associated with syndromes, in this case involving CHD, compiled in the Winter-Baraitser Dysmorphology Database (London Medical Databases, 2014). Following a study by Li *et al*. (2015) suggesting that genes involved with cilia development and cilia-transduced signalling play a major role in CHD, I decided to include these genes into my study. The mutated mouse orthologs of these genes were found to cause CHD (Li *et al.*, 2015). Some overlap between sets is to be expected. The overlap between categories pertaining to ToF and CHD in general are shown in Figure 23, where I have grouped ToF lists and CHD lists together. Most of the genes from the known ToF gene set are included in the IVA gene set. Comparing all the gene sets pertaining to ToF with the cilia gene set shows overlap only in 2 genes.

For each gene set, I calculated the odds ratio of ToF cases harbouring rare variants; first truncating followed by predicted deleterious, in candidate genes for ToF or CHD over NDD cases.

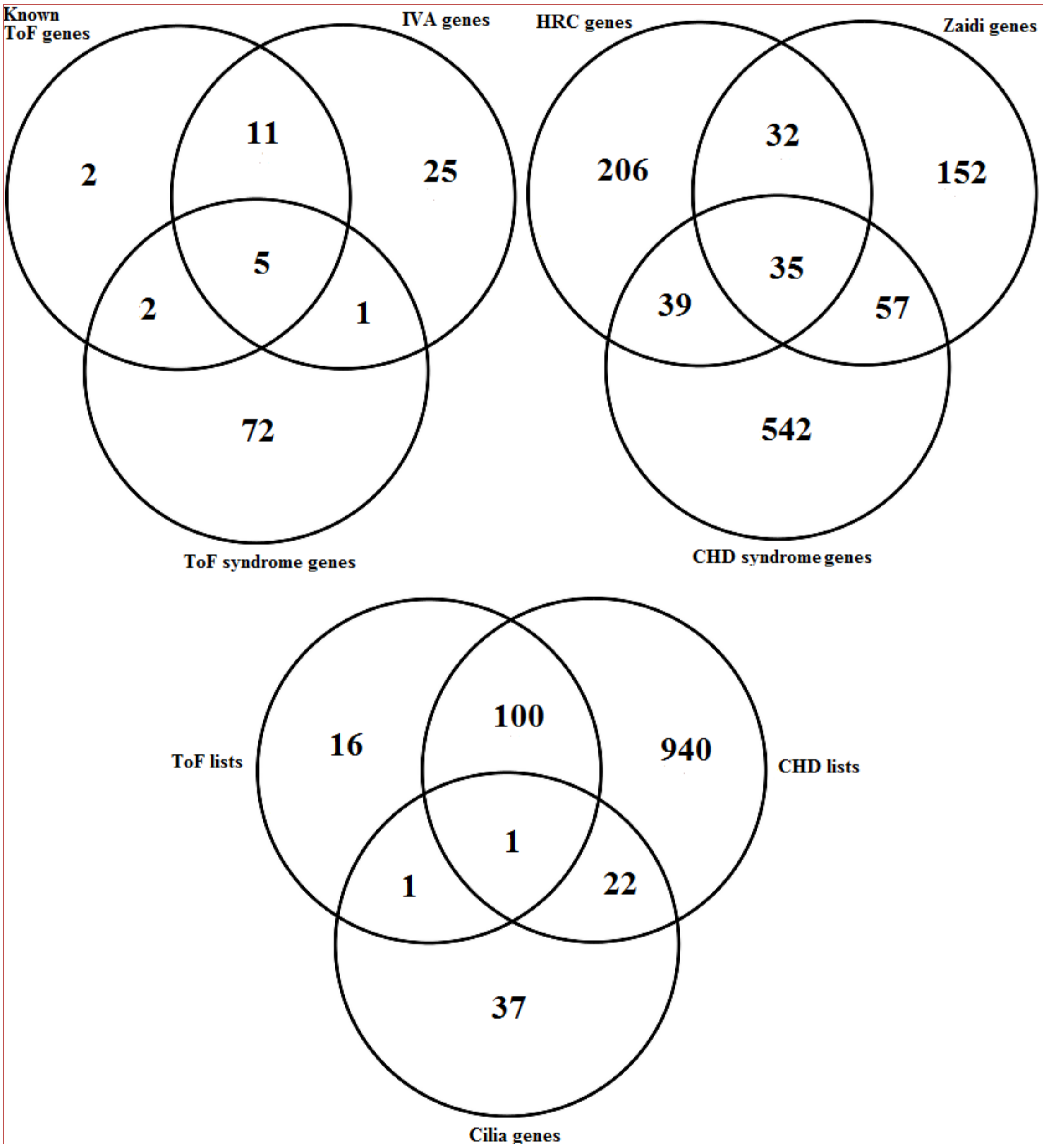

Figure 23. The overlap between the gene sets pertaining to ToF (top left) and between those pertaining to CHD (top right). Also shown is the overlap between all ToF gene sets and all CHD gene sets with the cilia gene set (bottom).

The next step in the WES ToF case study was to detect clustering in rare variants in the four categories delineated above. Variants contributing to disease are expected to cluster within specific gene or exons. The different categories provide a gradient of predicted effect against which the amount of clustering can be contextualised. For this, I implemented the Poisson approximation of The W_d statistic described by Lange (1997). Assuming that each position in the coding region of the exome is equally likely to carry a rare variant, the statistic estimates the probability of observing more than the expected number of rare variants for a given length of coding sequence (CDS) given the total number of variants in the exome and the total number of CDS and their lengths. I set the threshold for significance as $\alpha=0.05$. For this study, I define two sets of CDS: The coding portion of the gene and the exon. For the gene, coding regions were determined using known gene transcripts. The number of protein-coding genes was established to be 19,214, while the number of exons 233,785, from records in the Ensembl database (Cunningham *et al.*, 2015). These values were used to correct for multiple-testing. The identifiers and sizes of known protein-coding transcripts and specific exons were also obtained from the Ensembl database using BioMart to shape and download the query (Cunningham *et al.*, 2015). Ensembl provides transcripts which have different levels of support, signalled by a transcript support level (tsl) flag. For this study, I only used gene transcripts marked as tsl:1 which signifies the following: “all splice junctions of the transcript are supported by at least one non-suspect mRNA” (Cunningham *et al.*, 2015). I counted variants falling within a CDS using two different schemes. For the first scheme, if a variant at one particular position was seen in several different patients it was still only counted once. I refer to this first scheme as clustering by position. This ensured that any remaining non-rare variants or sequencing artefacts repeating across individuals did not lead to excessive clustering, while masking genuine clusters of rare variants. In a second scheme, I counted the actual number of rare variants in patients. I refer to this second scheme as clustering by variant. For comparison, I also looked at clustering within genes and exons in NDD cases. In order to compare clustering in ToF and NDD cases, I sampled variants in ToF cases 1000 times in each category, with each new sample containing the same number of variants as in NDD cases. This is an important step given the difference in the number of cases with ToF and NDD. I looked for any clustering in genes previously implicated in CHD using the gene sets described in Table 21.

Finally, I performed pathway analysis on the genes identified at different stages of this project. As in chapter 4, I started by using STRING v.10 (Jensen *et al.*, 2009), followed by EnrichNet (Glaab *et al.*, 2012). Using STRING, I looked at the interaction between the protein products of the genes found in each clustering category. I first looked for enrichment in interaction, followed by specific enrichment in gene ontology (GO) terms (Gene Ontology, 2004) and KEGG interaction pathways (Kanehisa and Goto, 2000). Using EnrichNet, I interrogated a number of pathway databases to uncover any over-representation of genes belonging to a certain pathway or process. The significance of overrepresentation is judged using two metrics: the level of overlap between query genes and genes in a pathway; given as a *q*-value, and the distance of query genes to genes in a pathway; given as an XD-score. The evidence gathered from different layers of enquiry was used to determine which genes could play a role in ToF.

5.4. Results

5.4.1. Sequence data presentation

My study originally focused on 867 ToF cases sequenced by MUGQIC and 500 NDD cases, sourced from various UK10K studies that were to be used for comparison. The mean depth of coverage for each set and the number of variants per case is given in Table 22 along with an overall estimation of the sensitivity and specificity using population data (Houniet *et al.*, 2015). The mean read depth per ToF case is nearly twice that of NDD cases. This likely stems from differences in sequencing protocol between case sets.

	ToF cases (n=867)	**NDD cases (n=500)**
Read depth per case (mean)	93.9	49.9
Number of Q30 variants per case (mean)	38589	39668.8
Estimated sensitivity (%)	95.9	93.3
Estimated specificity (%)	99.5	99.7

Table 22. Description of sequence data from ToF and NDD cases in target regions before the selection of cases with high specificity.

As an initial quality filtering step, I excluded cases that were in the lower specificity range, with an estimated specificity ≤99%. These are shown in Figure 24. Overall, 824 ToF cases and 490 NDD cases were estimated to have a specificity >99%. All other cases were not used for the remainder of the study.

Table 22 hints at some systematic difference between sets. This may be the result of variations in sequencing protocol which can affect the frequency with which a particular allele is observed in a set of cases. The variability in allele frequencies imputable to disease may consequently be masked by the variability introduced by different sequencing protocols. To determine whether the two sets are comparable, for each gene I compared the number of ToF and NDD cases with at least one high-confidence rare synonymous variant. I ended up with an individual count per gene for each set. The counts per gene in each set were compared using Fisher's exact test and the resulting distribution was compared to a chi square distribution. If the two sets have comparable variant counts, the resulting distribution should approximately follow a chi-square distribution. The distribution of my data did not follow the expected chi-square distribution, as shown in Figure 25. There was more variance in variant counts between the two sets than would be expected by chance alone. This manifests in Figure 25 as an upward curve. Figure 26 shows the same experiment over higher-confidence variants (Q60). Selecting variants from higher-confidence calls does not lead to a chi-square distribution. A comparison of allele frequencies in case sets, following the designs described in Section 5.2.2, is therefore likely to be compromised by sequencing bias.

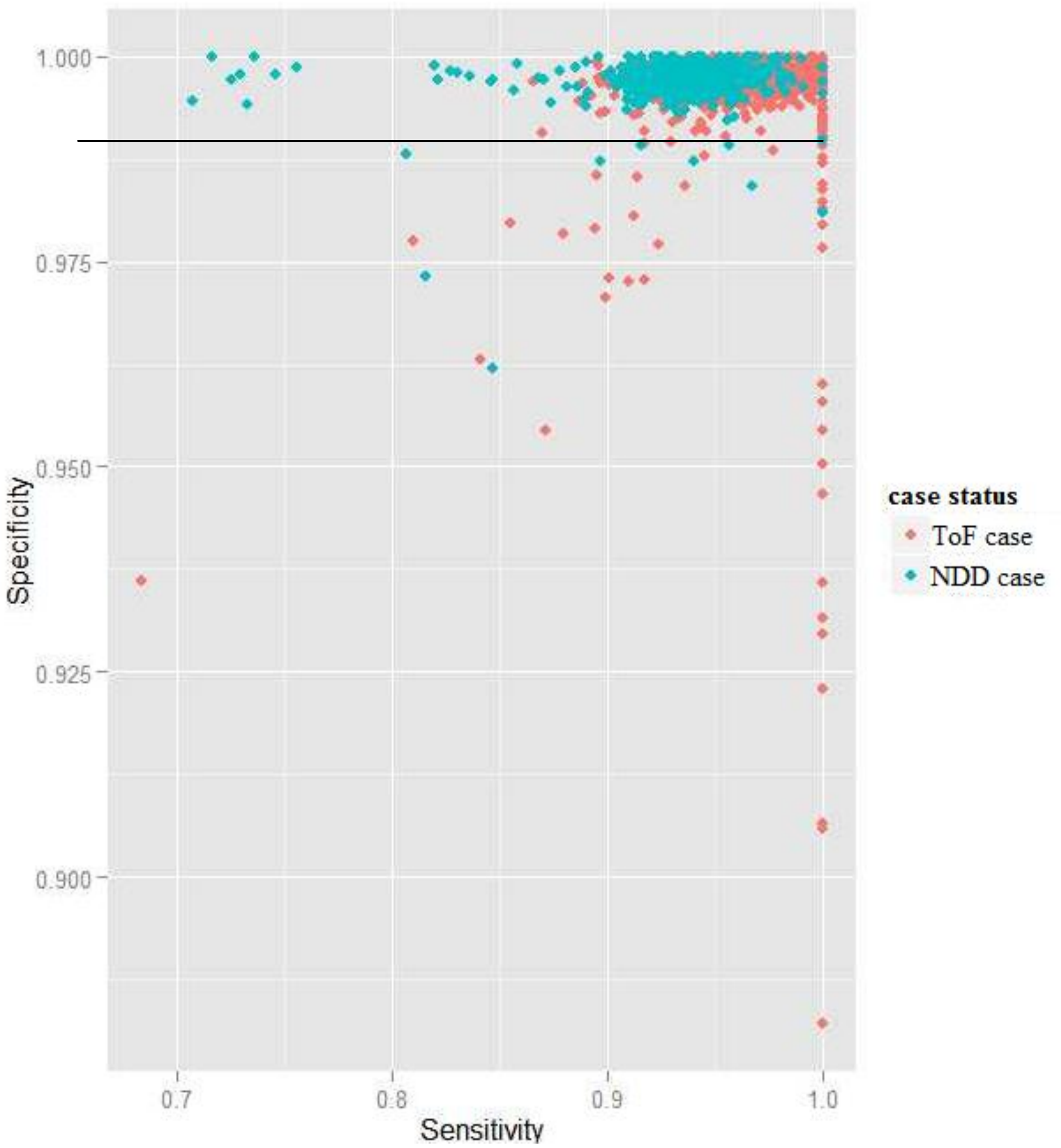

Figure 24. Estimates of sensitivity and specificity in 824 ToF cases and 490 NDD cases. The line represents a cut-off threshold of 99%. Cases below the threshold are excluded from the study. Data visualisation provided by Mauro Santibañez-Koref.

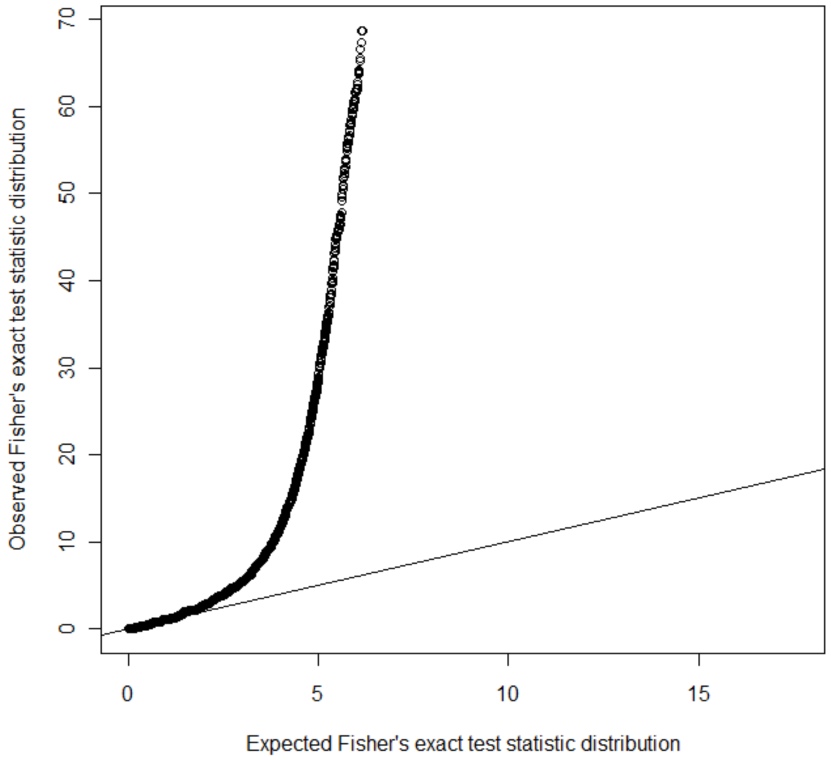

Figure 25. Distribution for each gene compared with a standard distribution of Fisher's exact test statistics. The distribution does not approximate the chi-square distribution.

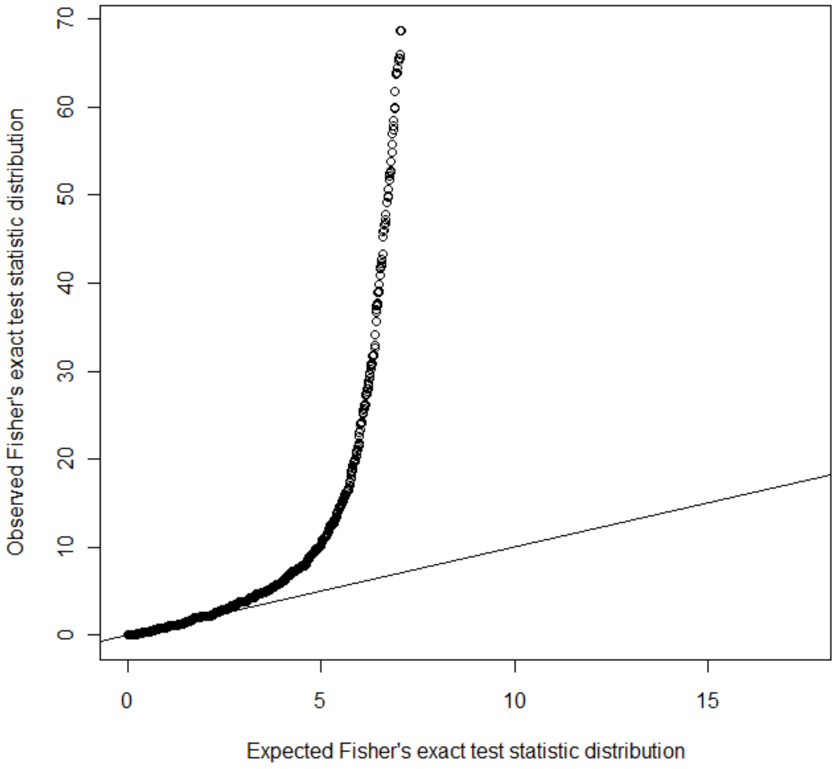

Figure 26. Distribution for each gene (variants Q60) compared with a standard distribution of Fisher's exact test statistics. Despite quality selection, the distribution does approximate the chi-square distribution.

Variant calls for each case were filtered to contain only high-quality rare variants. The mean count of these rare variants and the category they belong to are summarised in Table 23. There appears to be more rare SNVs of all categories per ToF case than per NDD case. The largest difference concerns truncating SNVs with around 1.5 times more variants per ToF case than per NDD case. In the categories that contain indels, the differences between ToF and NDD cases were more pronounced. There were around 3.8 more indels predicted deleterious per ToF case than per NDD case. The differences are not necessarily biologically meaningful and could also be the result of differences in sequencing protocols. Alternatively, it can be the result of different set sizes, with the rarer variants having a lower probability of being represented in a set of 490 cases over a set of 824 cases.

Category	**ToF cases (n=824)**		**NDD cases (n=490)**	
	SNVs	Indels	SNVs	Indels
Truncating	8.9	6.4	6	3.6
Predicted deleterious	47.1	7.5	39	2
Non-synonymous	105.9	0	81.1	0
Synonymous	79.1	0	56	0

Table 23. Mean count per patient of rare variants in each variant category. Categories are mutually exclusive.

5.4.2. Variants in genes implicated in congenital heart disease

Given known issues with indel calling using SAMtools, I decided to primarily focus on SNVs. I started by focusing on those rare SNVs categorised as truncating and therefore with the highest potential to be damaging. Do any of these rare truncating SNVs occur within genes previously associated with ToF or CHD? How many ToF cases are these rare SNVs found in? To answer these questions, I used the seven sets of genes previously associated with either ToF or a range of CHDs, described in Section 5.3. Answers are provided by the data presented in Table 24.

Gene set	Carrier frequency in ToF cohort (%)	Number of genes	Number of SNVs	Carrier frequency in NDD cases (%)	Number of genes	Number of SNVs
Known ToF	**0.8**	**5**	**7**	**0**	**0**	**0**
IVA	2.2	10	19	0.4	2	2
ToF syndrome	3.3	23	35	2.4	10	12
HRC	11	91	139	9.6	41	48
Zaidi	10.4	88	139	10.2	36	54
CHD syndrome	21.8	227	376	23.7	104	139
Cilia	5.9	30	67	4.7	15	23

Table 24. Rare truncating SNVs in ToF and NDD cases present in genes from CHD gene sets. For results with indels see Appendix, Table S3.

Of particular note is the fraction of ToF cases with truncating rare SNVs in known ToF genes. Rare truncating SNVs in known ToF genes were found in 0.8% of ToF cases. The truncating SNVs occurred across 5 ToF genes: *GATA6*, *JAG1*, *NKX2-6*, *NODAL* and *TBX1*. *NKX2-6* and *NODAL* both harboured two variants each. Allowing for genes more loosely implicated in ToF by using the set provided from IVA's data mining increased the fraction of ToF cases in which genes were found to 2.2%. Considering genes implicated in syndromes in which ToF has been observed instead of the isolated ToF genes from the previous two sets resulted in a carrier frequency in ToF patients of 3.3%. A little over 10% of ToF patients had at least one truncating variant in genes that belonged to the Zaidi or HRC gene sets. Rare truncating SNVs in genes implicated in syndromes with any CHD phenotype were found in 21.8% of cases. All three sets cover a larger array of genes than previous sets, particularly the latter. It is therefore to be expected that a higher fraction of ToF cases have truncating variants in genes that appear in these sets. Variants in genes belonging to the Cilia gene set were found in 5.9% of ToF patients. To establish whether the genes in each set are more prevalent in ToF cases than in some unrelated disease requires a point of comparison. I performed the same analysis on NDD cases, the results of which are shown in Table 24.

The most interesting result was that, in 490 individuals with NDD, not a single rare truncating SNV was found in a known ToF gene. Two individuals with NDD did have truncating variants in genes associated with ToF according to IVA. The two genes in question, *MYOM2* and *TCEB3,* originated from a single study by Grunert *et al.* (2014). *MYOM2* was found both in ToF cases and NDD cases. The differences between ToF and NDD patients for other gene sets were less striking. For example, a total of 2.4% NDD patients had at least one truncating variant in a gene previously implicated in syndromes that include ToF, compared to 3.3% in ToF patients. This result is altogether not unexpected given that some syndromes will involve both CHD and neurodevelopmental disorders. Patients with CHARGE syndrome for example; caused by mutations in *CHD7,* can present both ToF and mental retardation (Michelucci *et al.*, 2010). To establish whether genes in a particular gene set were significantly more prevalent in ToF cases than in NDD cases, I calculated an odds ratio for each set. The odds ratios are given in Table 25. The odds of truncating variants being found in genes associated with ToF; whether in the Known ToF or IVA gene set, were substantially higher in ToF cases than in NDD cases. However, this did not extend to genes implicated in syndromes of which ToF is a part. For other gene sets, Fisher's test did not reach statistical significance and the odds gravitated towards 1. Lower odds are to be expected when looking at CHD genes, many of which have yet to be shown to influence the risk of ToF. Conversely, many of them could be associated with neurodevelopmental disorders. This conclusion is particularly applicable to the CHD syndrome gene set. Additionally, the HRC gene set contains a subjective selection of CHD genes some of which could be only tenuously linked to CHD. Given that the two sets of cases were not well-matched, one must exercise great care in interpreting these results.

Gene set	Odds ratio (*p*-value)
Known ToF	**∞ (3.79E-2)**
IVA	**5.44 (6.76E-3)**
ToF syndrome	1.35 (2.49E-1)
HRC	1.17 (2.32E-1)
Zaidi	1.03 (4.86E-1)
CHD syndrome	0.9 (7.99E-1)
Cilia	1.28 (2.01E-1)

Table 25. The odds of ToF cases having truncating variants in relevant genes over NDD cases.

For both ToF and NDD patients, I looked at the distribution of truncating SNVs in genes from the HRC set. 80 of 91 genes harboured less than 3 variants in ToF cases. All 41 genes with SNVs in the NDD cases had less than 3 variants. Both ToF and NDD cases had truncating SNVs for 16 of these genes. Table 26 lists the few genes for which variants were found only in ToF patients and for which there were more than 3 truncating SNVs. For this gene set, *NOTCH1* was revealed to be the gene with the largest number of variants. *IGF2* presented an equal number of truncating variants, but these are all filtered out if stringent filtering is applied.

HRC genes	Number of truncating variants
IGF2, NOTCH1*	5
*EDN**	4
CHFR, CLTC, MESP1, PRKG1, SCN5A,VCL	3

Table 26. HRC set of genes for which there are at least 3 truncating SNVs in 824 ToF cases, but not in 490 NDD cases. *SNVs in gene completely removed by stringent filtering.

Having looked at rare truncating SNVs, I then turned to variants predicted as deleterious. It is worth reiterating at this stage that truncating SNVs were not included in this category.

The results for ToF and NDD cases are shown in Table 28. SNVs predicted to be deleterious in known ToF genes were found in 3.6% of ToF patients. It is important to note at this stage that a number of missense variants were found in known ToF genes in two gene sequencing studies carried out on a subset of the ToF cases being studied here (Griffin *et al.*, 2010; Topf *et al.*, 2014). *TBX1* and *HAND2* were among the genes sequenced in these two studies. In order to avoid bias from variant rediscovery, I carefully surveyed the variants found in my analysis and compared them to those found in the two gene sequencing studies. There was no overlap in the variants that were found. As shown in Table 23, per individual, the number of rare SNVs predicted to be deleterious is an order of magnitude larger than the number of rare truncating SNVs. This translated to a high carrier frequency for each gene set, whether ToF or NDD patients were being studied.

Table 27 provides the list of known ToF genes observed in ToF cases. Genes for which truncating SNVs were found all also harbour at least one SNV predicted to be deleterious in ToF cases. Notably, there were six SNVs fitting that description in *JAG1*.

Known ToF genes	**Number of variants predicted deleterious**
JAG1	6
GATA4	5
BVES, *TBX5*, *ZFPM2*	3
FOXH1, *GATA5*, *GATA6*, *NODAL*	2
GJA5, *HAND2*, *NKX2-6*, *TBX1*, *TDGF1*	1

Table 27. Known ToF genes for which there are variants predicted deleterious in 824 ToF cases.

NDD patients also presented variants in known ToF genes. Given that the damaging potential of these variants is more uncertain; as it relies on pathogenicity prediction tools, this result was not entirely surprising. Five genes for which SNVs were found in ToF patients also harboured SNVs in NDD patients: *BVES*, *FOXH1*, *JAG1*, *NKX2*-6 and *TBX5*. Three variants were found in *JAG1* in NDD patients for example. The contrasts between ToF and NDD patients observed when studying rare truncating variants in known ToF genes did not carry over to rare variants predicted as deleterious. However, the odds ratios did reveal some differences, as shown in Table 29.

Gene sets	Carrier frequency in ToF cases (%)	Number of genes	Number of SNVs	Carrier frequency in NDD cases (%)	Number of genes	Number of SNVs
Known ToF	3.6	14	33	1.8	5	9
IVA	9.2	32	85	8.2	16	42
ToF syndrome	27.8	54	289	17.6	42	93
HRC	65	229	995	58.6	172	449
Zaidi	59.5	184	877	54.7	143	396
CHD syndrome	89.2	469	2255	87.6	391	1106
Cilia	30.1	44	312	33.5	43	205

Table 28. Rare SNVs predicted to be deleterious in ToF and NDD cases that are present in genes from CHD gene sets. For results with indels see Appendix, Table S4.

ToF patient presented more than twice the odds of having a variant predicted to be deleterious in a known ToF gene as an NDD patient. Given the increased number of SNVs included in this category, statistical significance was reached for several gene sets. This was the case for the HRC and the ToF syndrome gene sets. The latter gene set presented an odds ratio of 1.8 in favour of ToF patients. Whether statistically significant or otherwise, other gene sets had odds ratios close to 1.

Gene sets	**Odds ratio (*p*-value)**
Known ToF	**2.02 (4.18E-2)**
IVA	1.14 (2.91E-1)
ToF syndrome	**1.8 (1.33E-5)**
HRC	1.32 (1.12E-2)
Zaidi	1.22 (5.11E-2)
CHD syndrome	1.17(2.06E-1)
Cilia	0.86 (9.09E-1)

Table 29. The odds of ToF cases having variants predicted as deleterious in relevant genes over NDD cases.

To determine if SNVs in specific genes of the ToF syndrome set might be responsible for the observed odds ratio, I looked at gene-specific odds ratios and their distribution in ToF and NDD cases. Genes for which variants predicted to be deleterious were found largely overlapped in ToF and NDD cases. Among the 10 genes with the most SNVs in ToF cases, only *PDX1* was not represented in NDD cases. However, on further inspection, many of the 13 individuals with variants in *PDX1* were found to share one of two recurring variants, leaving only two unique variants. To compare variants in ToF and NDD cases, I therefore chose to focus only on the proportion of individuals with unique variants. Furthermore, for each gene only one variant was counted per individual. The results of this analysis are shown in Table 30. The odds of a variant being observed in *NOTCH1* in a ToF patient were 5 time higher than in an NDD patients. Results for other genes did not reach statistical significance.

ToF syndrome genes	SNVs pred. deleterious in ToF patients	SNVs pred. deleterious in NDD patients	Odds ratio (*p*-value)
NOTCH1	**25**	**3**	**5.07 (1.63E-3)**
DOCK6	15	3	3.01 (5.22E-2)
ATP7A	9	1	5.39 (6.46E-2)
CHD7	9	3	1.79 (2.86E-1)
ACE	9	7	0.76 (7.89E-1)
GLI3	6	3	1.19 (5.51E-1)
EHMT1	8	1	4.79 (9.45E-2)
BOC	8	3	1.59 (3.63E-1)

Table 30. Genes from the ToF syndrome set with the most SNVs predicted to be deleterious after recurring variants have been excluded.

Whether the focus is on truncating variants or variants predicted to be deleterious the odds of finding rare variants in known ToF genes in ToF cases rather than NDD cases were consistently higher. Additionally, there were slightly higher odds of finding variants predicted as deleterious in genes from the ToF syndrome gene set in ToF patients than in NDD patients. The overabundance of *NOTCH1* carried part of that trend.

5.4.3. Identification of variants clustering in genes and exons

Results like those shown for *NOTCH1* in Table 30 suggest that rare variants are clustered within certain genes in ToF patients. Following the assumption that each position in the coding sequence (CDS) of the exome is equally likely to carry a variant, clustering then refers to a higher concentration of variants than expected by chance within a stretch of CDS given the overall variant rate. If ToF patients harbour variants that are disease-causative, some level of clustering across corresponding genes should be discernible in the disease-specific population. The second step in this WES study consisted in identifying such clusters in ToF cases and sorting them by gene. Genes for which clusters were found in ToF and NDD cases were excluded as not disease-specific. It is important to stress that this does not constitute a comparison between ToF and NDD cases, as the two sets were not well-matched, as shown in Section 5.4.1. The NDD cases here provided a filter for clusters that were not specific to ToF cases. Genes were also excluded if clustering was also found in the Synonymous category in ToF cases given that synonymous variants are unlikely to lead to be disease-causative.

As outlined in Section 5.3, I started by looking at clustering of SNVs by position. These results are shown in Table 31 for every category of rare variant at gene and exon level.

The number of genes in which clusters occur at both levels was also determined. Clusters were identified in a total of 144 genes. Among these genes, 13 were identified both in ToF and NDD cases. Crucially, none of these overlapping genes were found in the Truncating category. This is significant given that the category concentrates the most likely highly-deleterious variants. Looking at the overlap between the categories in ToF cases showed 15 genes present in at least two categories. 7 of the 15 genes overlapping categories in ToF cases were also present in NDD cases. The recurrence of these genes is likely to be the product of their high tolerance to mutation, leading to false cluster detections. Both the Non-synonymous and Synonymous categories, which are expected not to concentrate disease-causing variants presented many clusters.

CDS type	ToF cases (n=824)	NDD cases (n=490)
Truncating	6,802	2,623
Gene	3	1
Exon	7	6
Both	1	0
Predicted deleterious	34,065	16,454
Gene	30	11
Exon	19	8
Both	4	0
Non-synonymous	73,043	34,627
Gene	37	9
Exon	19	9
Both	7	0
Synonymous	53,767	24,159
Gene	18	4
Exon	8	6
Both	2	0

Table 31. Number of genes that showed evidence of clustering (by position) at gene and exon levels. The total number of positions in each category in ToF and NDD cases is given in the header of each section. Only SNVs were considered here.

Despite not being able to compared ToF and NDD cases directly, I was able to compare their rates of clustering. To do this, the same number of positions had to be studied in both. As shown in Table 31, the number of positions studied for ToF cases is at least twice that of NDD cases, due not only to the higher number of ToF cases, but also differences in sequencing. In order to get an approximation of the number of clusters that would be observed in ToF cases given the same number of positions as in NDD cases, I sampled the positions in ToF cases 1000 times to match the number of positions in NDD cases. For example, for the truncating category, I created 1000 samples from ToF cases covering 2,623 positions randomly selected from the original 6,802. For each of these lists, I uncovered clusters around gene and exons and produced an average number of clusters. The results of this process are shown in Table 32. With some exceptions, ToF and NDD cases showed similar numbers of clusters. If disease-causative variants were responsible for most ToF cases, more clustering in ToF cases should have been observed in the Truncating and Predicted Deleterious category. No clear indication of this was found.

CDS type	**NDD cases (n=490)**	**1000 ToF cases samples (from n=824)**	
		Average	Fraction with more clusters than NDD cases
Truncating		2,623	
Gene	1	1.7	0.51
Exon	6	7.5	0.64
Predicted deleterious		16,454	
Gene	11	9.4	0.23
Exon	8	8.8	0.53
Non-synonymous		34,627	
Gene	9	11.4	0.76
Exon	9	6.5	0.07
Synonymous		24,159	
Gene	4	4.9	0.57
Exon	6	4.1	0.04

Table 32. Average number of genes which showed evidence of clustering (by position) at gene and exon levels after sampling of positions in ToF cases (NDD cases are shown for comparison). The fraction of samples that produced a higher number of clusters than observed in NDD cases is also indicated. Only SNVs were considered.

Clusters of truncating SNVs were found across 9 genes, most of these identified within a single exon. Most clusters were found to include no more than 3 truncating SNVs. The exceptions were clusters in genes *CCDC102B* and *FLT4,* with cluster sizes in the gene of 5 and 8 SNVs respectively; where cluster size refers to the number of SNVs in the cluster. The clusters were specific to the Truncating category, the only exception being a cluster of 8 SNVs in the non-synonymous category across *CCDC102B.* Table 33 describes those genes in which clusters were found. The clusters are not affected if stringent filtering is applied.

Gene	Clustering level	Cluster size (*p*-value)
A2M	Exon	2 (3.72E-2)
CCDC102B	Gene/ Exon	5 (3.57E-2) / 3 (2.65E-2)
CTPS2	Exon	3 (2.74E-3)
FLT4	Gene	8 (1.28E-2)
PHF10	Exon	3 (9.16E-3)
RAD9B	Exon	3 (4.07E-3)
SHPK	Exon	3 (2.98E-2)
TAC3	Gene	3 (4.36E-2)
TNPO1	Exon	3 (2.98E-3)

Table 33. Size and *p*-value of clusters (by position) of truncating SNVs in ToF patients. The clusters of truncating variant in NDD patients do not occur in the same genes.

There were considerably more clusters in the Predicted Deleterious category. In total, clusters were found in 48 genes. Among these, 4 genes were also the site of clustering in NDD cases: *ATG2A*, *PCDH9*, *PRRC2C* and *ZNF423*. SNVs in the Synonymous category in ToF cases were also found to cluster in *PRRC2C* and *ZNF423* and *APC2*. The same genes were found in the Non-synonymous category as well, with the addition of *ATG2A*. The abundance of SNVs in these genes across multiple categories and, more importantly, across cohorts suggested a high gene mutation rate independent of ToF. Applying stringent filtering to the list leads to the further exclusion of 5 genes: *ADAMTS17*, *MEF2D*, *SLC26A1*, *STX6* and *VCL*. Table 34 lists the 10 genes with the most significant clustering relative to gene or exon size; excluding the 9 genes just mentioned. The full list can be found in the Appendix as Table S6. The list includes *NOTCH1* which was identified in Section 5.4.2 as having an overabundance of variants in ToF patients compared to NDD patients. The presence of a cluster in *NOTCH1* in ToF cases, but not in NDD cases thus consolidates previous results. The largest clusters were found in *NEB* and *HMCN1* with a variant count larger than 50. Unlike clusters in the previous category, a majority of clusters were found spanning several exons rather than being concentrated within a single one.

Gene	**Clustering level**	**Cluster size (*p*-value)**
HMCN1	Gene	56 (4.37E-7)
SERBP1	Gene/ Exon	14 (6.18E-7) / 9 (1.92E-5)
MYO7A	Gene	30 (6.36E-6)
NAV2	Gene	32 (6.38E-6)
GLDC	Gene	20 (9.11E-6)
NEB	Gene	58 (2.83E-5)
NELL1	Gene	17 (4.80E-5)
NOTCH1	Gene	31 (7.89E-5)
SNCAIP	Gene	6 (8.95E-5) / 4 (2.99E-2)
PTPN5	Exon	6 (1.35E-4)

Table 34. Size and *p*-value of the 10 most significant clusters (by position) of SNVs predicted to be deleterious in ToF patients, excluding those with an equivalent in NDD cases and the Synonymous SNV category. The full list is in the appendix in Table S6.

After looking at clusters by position, I looked at clusters formed by variants counted for each occurrence; which I refer to here as by variant. The number of clusters in each category for ToF and NDD cases is shown in Table 35. Here again I considered only SNVs. This approach to clustering led to a sharp increase in each category both in ToF and NDD cases when compared to results from clustering by position. This suggests, despite filtering for rare variants, that many variants shared by multiple unrelated individuals remained in the data. The application of stringent filtering supports this view as large numbers of clusters from the first three categories were discarded as a result. This was particularly true of clusters in exons for both the Predicted Deleterious and Non-synonymous categories with 79 and 126 clusters removed after stringent filtering. The results after stringent filtering are shown in the Appendix in Table S5. A total of 51 genes were found in at least two categories, 14 of which were also found in NDD cases. However, there was once more little overlap between the Truncating category and other categories. *CCDC102B* was once again shared by Truncating and Non-synonymous categories. *GCKR* and *TPO* were shared with the Synonymous category. *FLT4*, previously found to harbour significant clustering in the Truncating category for ToF cases, was also found in the same category for NDD cases. Crucially however, the clustering was limited to exon 2 and involved the same exact variant in 4 different NDD patients. By contrast, the cluster in ToF cases involved 8 distinct rare variants. Therefore, the presence of a cluster in *FLT4* in NDD patients does not affect the importance attributed to the cluster in ToF patients.

CDS type	TOF cases (n=824)	NDD case (n=490)
Truncating	7,326	2,964
Gene	34	34
Exon	60	81
Both	24	25
Predicted deleterious	38,826	19,105
Gene	80	31
Exon	186	159
Both	43	18
Non-synonymous	87,247	39,763
Gene	126	29
Exon	222	54
Both	67	9
Synonymous	65,177	27,453
Gene	83	13
Exon	255	58
Both	50	7

Table 35. Number of genes that showed evidence of clustering (by variant) at gene and exon levels. The total number of variants in ToF and NDD cases is given for each category. Only SNVs were considered.

I once again compared ToF and NDD cases through sampling. The results are shown in Table 36. The numbers of genes with clusters in each category varies widely between ToF and NDD cases. This is true of categories where little variation was expected such as the Synonymous category. There were almost systematically less clusters in the samples of ToF cases for the Truncating and Predicted Deleterious categories compared to NDD cases. The opposite tendency was observed in the remaining two categories. One possible explanation for this observation is that it represents disease-specific clustering in the NDD cases. However, it is difficult to for a conclusion in this case given that ToF and NDD cases were not well-matched. The excess clustering in NDD cases could also be due to a higher number of artefacts in that set of cases over ToF cases.

CDS type	NDD cases (n=490)	1000 TOF cases samples (from n=824)	
		Average	Fraction with more clusters than NDD cases
Truncating	2,964		
Gene	34	11.1	0
Exon	81	24.4	0
Predicted deleterious	19,105		
Gene	31	25.5	0.07
Exon	159	68.2	0
Non-synonymous	39,763		
Gene	29	38.8	0.99
Exon	54	74.1	1
Synonymous	27,453		
Gene	13	20.8	0.97
Exon	58	66.4	0.87

Table 36. Average number of genes which showed evidence of clustering (by variant) at gene and exon levels after sampling of positions in ToF cases (NDD cases are shown for comparison). The fraction of samples that produced a higher number of clusters than observed in NDD cases is also indicated. Only SNVs were considered.

In the following sections, I have therefore decided to focus on results obtained from the study of clustering when SNVs were counted by position rather than by variant.

5.4.4. Clustering in genes implicated in CHD

Having identified clusters in genes and exons, some fraction of which could be hypothetically explained by genes influencing the risk of ToF, I then compared the sets of CHD genes to all clusters for ToF case. Were any of the genes in which clusters were found already implicated in ToF or CHD more generally? The results are given in Table 37, with a gene by gene break down for each category in the Appendix, Table S7. Table 38 provides the list of genes for which clusters were found in the Predicted Deleterious category for ToF cases, but not in the Synonymous category or in NDD cases.

Gene sets	Truncating	Predicted Deleterious	Non-Synonymous	Synonymous
Known ToF	0	0	0	0
IVA	0	0	0	0
ToF syndrome	0	2	0	0
HRC	0	4	2	1
Zaidi	0	2	1	0
CHD syndrome	0	6	4	4
Cilia	0	1	0	0

Table 37. Number of genes from CHD gene sets that harbour clusters (by position) for ToF cases. There is some overlap between genes found in each gene set. For list of genes see appendix, Table S7.

None of the genes included in ToF and IVA sets presented clustering. Clusters in two genes of the ToF syndrome set were found in the Predicted Deleterious category, *NOTCH1* and *PORCN*; *NOTCH1* is also shared by the HRC, Zaidi and CHD syndrome sets. As in sections above, the cluster in *NOTCH1* stood out with 31 SNVs.

Gene	**Gene set**	**Clustering level**	**Cluster size (*p*-value)**
*ADAMTS17**	CHD syndrome	Gene	17 (4.15E-3)
CC2D2A	CHD syndrome, Cilia	Exon	4 (4.07E-2)
COL3A1	CHD syndrome	Gene	19 (1.62E-2)
*MEF2D**	HRC	Gene	6 (2.00E-2)
MYOM1	HRC	Exon	5 (3.19E-3)
NEK2	Zaidi	Gene	10 (2.09E-2)
NOTCH1	ToF/CHD syndrome, HRC, Zaidi	**Gene**	**31 (7.89E-5)**
PORCN	ToF/CHD syndrome	Exon	4 (4.72E-2)
*VCL**	HRC	Gene	13 (4.74E-2)

Table 38. Genes present in CHD gene sets for which there are clusters (by position) in the Predicted Deleterious category only for ToF cases. Also excluded are genes in the Synonymous category. *Gene no longer in list if stringent filtering is applied.

5.4.5. Pathway analysis.

I started by submitting genes in which SNVs cluster to STRING v.10 (Jensen *et al.*, 2009). I then ran these same genes through EnrichNet (Glaab *et al.*, 2012). STRING required at least 10 genes in order to calculate interaction enrichment. I therefore combined the Truncating and Predicted Deleterious categories for this step, producing a total list of 54 genes. The potential protein-protein interactions identified by STRING are shown in Figure 27. STRING determined the set of genes not to be enriched in interactions (*p*-value=6.60E-1). A total of 4 interaction networks were found, including one with 4 interacting proteins. The networks included genes found in the HRC gene set *MEF2D*, *NOTCH1* and *VCL*. On closer inspection, not all protein-protein interactions appeared convincing. Evidence of protein-protein interactions for the protein products in the network containing 4 proteins; NMNAT1, LARS, CTPS2, PUS7, relied entirely on evidence from homologous genes in other species. As a result, the highest combined score (CS) for that network was 0.479, between NMNAT1 and LARS. The predicted interactions between NEB and VCL (CS=0.472) and between COL3A1 and ZNF423 (CS=0.437) relied solely on the co-mention of proteins and genes in publications, a fairly vague criterion to base interaction on. By contrast, the putative interactions between

MEF2D and NOTCH1 (CS=0.764) and between NOTCH1 and FLT4 (CS=0.946) showed stronger interactions, with varied types of evidence. In addition to co-mention of NOTCH1 and FLT4 in 50 publications, homologs in other species were found to interact in experimental and biochemical interactions. This result is particularly interesting given that the network included both *NOTCH1* and *FLT4*, genes singled out in previous sections.

For comparison, I performed the same step for NDD cases. This produced a list of 26 genes. STRING determined that there was no enrichment in interactions in this list either (*p*-value=1.16E-1). A single interaction network was found, including the proteins of genes *GRIK*, *DLG3* and *TANC1*. Interestingly, *GRIK* is associated with depersonalisation disorder; a type of psychosis, and *DLG3* with mental retardation, suggesting this network could be of biological significance for NDD cases. However, the combined score for their predicted interaction was only 0.476. Running either of these lists of genes through EnrichNet did not return any significant enrichment for known pathways or processes. The genes *GRIK* and *DLG3* were found to belong to a group of ionotropic activity of kainite receptors, but this was only significant in terms of network distance distribution (XD-score=1.8) and not overlap (*q*-value=6.7E-2).

As an additional test, I submitted all genes with at least one unique rare truncating SNV in ToF cases, and then in NDD cases, to EnrichNet. This resulting in a gene set of 4,971 genes in ToF cases and 2,309 genes in NDD cases, a large fraction of the total of human genes. Both resulting gene sets were enriched for ABC transporters (ToF cases: XD-score=3.2 and *q*-value=2.2E-4; NDD cases: 2.7 and *q*-value=2.2E-5) according to KEGG, while the gene set for ToF cases was also enriched for genes involved in glycosaminoglycan degradation (XD-score=3.169 and q-value=4E-2).

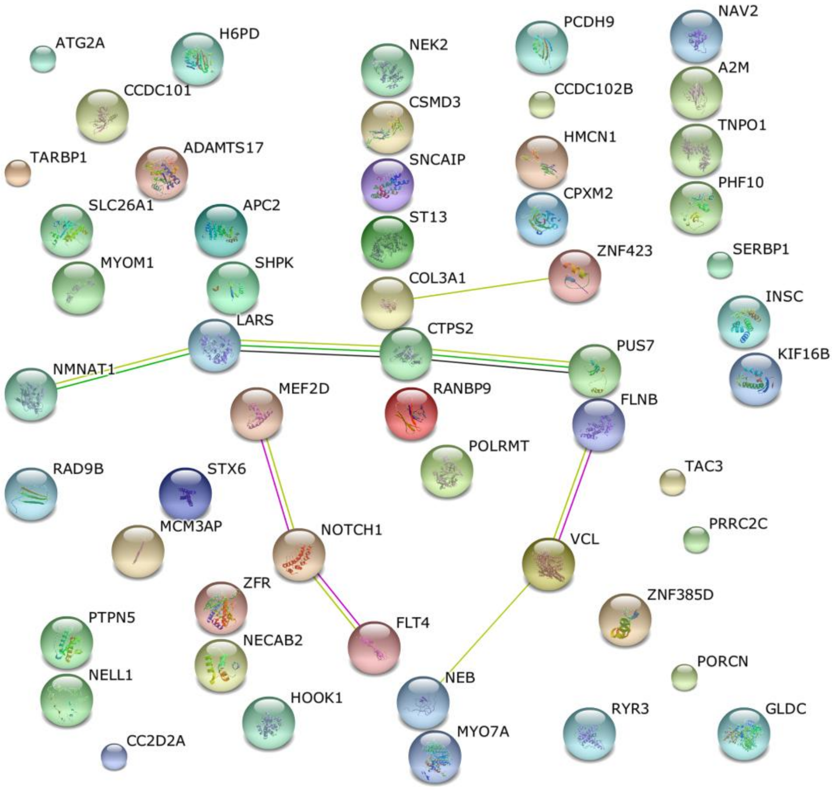

Figure 27. Protein-protein interactions for genes in which clusters of SNVs in Truncating and Predicted Deleterious categories were found.

While the functional link found between NOTCH1 and FLT4 could prove to be significant if FLT4 is confirmed to play a role in ToF, the STRING analysis did not lead to any new conclusions regarding potential pathways involved in ToF. The results from EnrichNet pathway analyses were similarly inconclusive.

5.5. Discussion

The work presented in this chapter provides both evidence of overrepresentation of known and potential ToF genes in ToF patients compared with NDD patients. In this Section I will review the evidence gathered for different genes.

5.5.1. Re-discovery of known ToF genes

In Section 5.4.2, I reported that rare truncating SNVs in known ToF genes; genes previously implicated in ToF via gene sequencing studies (Table 18), were found in 0.8% of ToF patients while being completely absent from NDD patients. This involved 5 of 20 known ToF genes with *NKX2-5* and *NODAL* each harbouring two SNVs each. 3.6% of ToF patients were found to harbour SNVs predicted as deleterious in 14 of 20 known ToF genes. Many of these SNVs were found in *JAG1* and *GATA4*. However, rare SNVs were also found in 1.8% of NDD patients. Both sets of patients had variants in the same set of 5 genes. One particular striking example is *JAG1*, for which 6 SNVs were found in ToF patients and 3 in NDD patients. On the other hand, there were also genes with rare SNVs only in ToF patients, such as *GATA4* with a total of 5 SNVs. ToF patients were shown to have twice the odds of having a rare variant predicted deleterious in a known ToF gene than NDD patients. But what can explain the 1.8% of NDD patients with SNVs predicted to be deleterious in known ToF genes?

There are two main reasons such results should be expected. The first reason is that a number of the variants detected in either set of patients may have been determined as deleterious when they are in fact benign. Inferring disease-causing potential for a missense variant requires the use of in-silico functional prediction tools with limited accuracy (Schwarz *et al.*, 2010). Given that these SNVs would have no influence on the incidence of ToF they would be unlikely to be found concentrated in either cohort. However, we would still expect to be able to detect the effect of those SNVs correctly predicted as deleterious as is likely to be the case in my data. It is also worth noting that other deleterious variants might be predicted as benign by at least one functional prediction tool, in which case they would fall in the Non-synonymous category and not be counted. The second reason we should expect such results is that, among the truly deleterious SNVs, there might be some that can contribute to neurodevelopmental disorders as well as ToF. Rare SNVs found in *JAG1* and predicted as deleterious were found in both ToF and NDD patients. Mutations in *JAG1* have been implicated in Allagille Syndrome which in turn have been known to include neurodevelopmental defects such as mental retardation in some rare cases (Rauch *et al.*, 2006). It could therefore be hypothesised that SNVs in a gene like *JAG1* would contribute to ToF in one set of patients and neurodevelopmental disorders in others. Despite this potential overlap, rare deleterious SNVs in ToF genes should be more consistently implicated in isolated ToF and therefore present in larger proportions among ToF patients over NDD patients.

This final observation is more difficult to apply to genes in which ToF is part of a syndrome. Not only do these syndromes not always involve ToF, they can involve many types of neurodevelopmental disorders. As was discussed in Section 5.2.1, 22q11.2 deletion syndrome is a recurrent cause for ToF. It has also been associated to psychosis and ASD, disorders represented in the NDD cases in this study (Wu *et al.*, 2014). Despite this, the data appears to indicate higher odds of ToF patients having rare variants predicted deleterious in genes for syndromes involving ToF (see Table 29). A gene by gene analysis revealed several genes to be more frequent in ToF cases rather than NDD cases although only the overabundance of SNVs in *NOTCH1* produced statistically significant results. Given that the ToF and NDD cases are not well-matched set, these results can only be interpreted as indicative of a potential role for *NOTCH1* in ToF, with additional evidence needed to complement this finding. It is worth noting that truncating SNVs in *NOTCH1* were also found in 5 ToF patients and not in NDD patients. *NOTCH1* is catalogued as a cause of Adams-Oliver syndrome; the main feature of which are skin and limb abnormalities, in the Winter-Baraister Dysmorphology Database (London Medical Databases, 2014). CHD is observed in 20% of cases of Adams-Oliver syndrome and one possible CHD subtype is ToF (Stittrich *et al.*, 2014). Incidentally, *NOTCH1* is not the only gene from this set to be involved in Adams-Oliver syndrome. *DOCK6*, with an odds ratio near statistical significance, has also been implicated in the syndrome. One paper by Wessels and Willems (2010) claims that *NOTCH1* has been implicated in non-syndromic ToF. This appears to be based the identification of a *NOTCH1* mutation in a family with aortic valve disease (Garg *et al.*, 2005). The only individual in the pedigree with ToF was not evaluated as deceased at the time, but the presence of a *NOTCH1* mutation in other individuals with CHD in the same pedigree does provide some evidence that *NOTCH1* is implicated in isolated ToF. There is also additional evidence for a role for *NOTCH1* in ToF from the study of CNVs. A study of 34 infants with isolated ToF; which do not include cases of 22q11 deletion syndrome, by Bittel *et al*. (2014) revealed two patients with CNVs encompassing the *NOTCH1* gene. In another study of CNVs in 114 ToF patients, Greenway *et al*. (2009) identified a single individual with 6 CNVs coinciding with *NOTCH1* and *JAG1* gene regions. Rare truncating SNVs in *NOTCH1* were found in 0.6% of ToF patients, while SNVs predicted as deleterious were found in 4.1%, a proportion larger than for the list of 20 known ToF genes. Additionally, in Section 5.4.5, *NOTCH1* was found to have 10 possible functional links with known TOF genes, particularly *JAG1*, another gene from the Notch-signalling pathway.

Results from the IVA gene set indicate that the odds of observing a rare truncating SNV in a gene of the IVA set were more than five times higher for ToF patients than NDD patients. However, this difference between ToF and NDD patients disappeared when SNVs predicted as deleterious were considered instead. What could explain such a strong disparity between the results from the set of known ToF genes and IVA genes? As shown in Figure 23, most genes from the known ToF set are accounted for in the IVA gene set. Of the remaining 25 genes identified by IVA, a total of 16 were selected for inclusion based on a single paper by Grunert *et al.* (2014). This implies that any shortcomings with the methodology of that particular study could have a large effect on the observed results for that particular gene set. The fact that the odds did not favour either ToF or NDD patients when rare variants predicted to be deleterious were considered suggests that the odds ratio observed for truncating SNVs could be largely led by known ToF genes rather than a feature of the IVA gene set.

What results obtained using ToF-oriented gene sets broadly suggest is that the genes previously implicated in ToF; whether by gene sequencing studies or through other methods, can only account for a small fraction of ToF cases. It cannot be determined which fraction of patients have truly deleterious variants in known ToF genes, but it cannot exceed 4.4% of patients.

The final results to consider are those that arise from the CHD gene sets, mainly the HRC and Zaidi sets. The genes in these sets do not appear to be harbouring proportionally more variants in either ToF or NDD patients. In Section 5.4.2, I pointed to the size of these sets compared with ToF specific gene sets and problems with the methodology of selection as potential reasons for these results. A case can be made that this provides further evidence that distinct CHD subtypes should be studied separately. Many deleterious gene variants will be specific to particular types of CHD. Selecting rare variants in ToF patients previously associated with CHD in general might therefore not yield meaningful results. In the case of genes from the study by Zaidi *et al*. (2013), this could be indicative of histone-modifying genes playing very little role in ToF specifically. This was the case for this study. The combination of ciliopathies and CHD in the Cilia gene set did not produce any distinct signal either in ToF patients either. This suggests that if cilia genes do influence the risk of CHD, this is not particularly applicable to ToF. Admittedly, many of these genes, when mutated, were accompanied by heterotaxy, a phenotype not commonly associated with ToF (Li *et al.*, 2015) .

In Section 5.4.2, I presented the identification of a number of new SNVs in known ToF genes as well as SNVs in *NOTCH1* which is occasionally described in the literature as a cause for ToF, although not systematically. The existence of these SNVs remains to be validated by Sanger sequencing in future work. Several genes not previously implicated

ToF, but involved in CHD which have appeared to be more abundant in ToF patients could play a role in the disease. However, it is not possible to reach a conclusion for these genes without an adequate comparator set. Additional comparisons of ToF cases with a well-matched reference set should provide additional evidence for some of these gene candidates.

5.5.2. Genes potentially involved in isolated ToF from clustering

The number of clusters varied significantly depending on the approach taken. The number of clusters varied strongly depending on whether clusters were determined by position or by variant. Determining clusters by counting each variant regardless of whether it had appeared in other individuals already led to a sharp increase in the number of clusters. Through sampling of variants in ToF patients, I was able to compare the proportional number of clusters in ToF and NDD cases (see Table 36). Where clusters had been obtained by counting each occurrence of a variant instead of each variant position once, NDD cases had more clusters of the Truncating and Predicted Deleterious category while ToF cases had more clusters of the Non-synonymous and Synonymous category. By comparison, counting variants by position led to ToF and NDD cases with similar numbers of clusters as evidenced in sampling results (see Table 32). This suggested that there was no excess of clustering in ToF cases compared to NDD cases. Assuming that ToF is largely brought on by single highly-penetrant rare variants while NDDs are the product of several interacting common and rare variants, one would expect to see more clustering in ToF patients than in NDD patients. This in turn suggests that highly-penetrant rare variants do not play a major role in ToF.

The first observation that can be made about genes for which there is evidence of truncating SNV clustering in ToF cases is that many of these clusters were extremely small and many resided near the threshold of statistical significance (see Table 33). One example was the cluster detected in *A2M* in exon 35. It is unlikely that a cluster that includes only 2 SNVs will truly be significant. This finding most likely reflects the small size of exon 35. Only two genes were the site of clustering involving more than 3 variants, *CCDC102B* and *FLT4*, with clusters of 5 and 8 respectively. Clustering was also found in *CCDC102B* in the Non-synonymous category. Both of these genes thus appeared to have a genuine overabundance of truncating SNVs in ToF patients. In the case of *FLT4*, this fact must be balanced with the discovery of a cluster in NDD cases as well when clusters were counted by variant. However as described in Section 5.4.3, this was the result of a single variant identical in four NDD patients and thus not the product of a very rare variant.

A number of other details about the gene suggest a potential role. *FLT4*; also previously described as *VEGFR-3*, encodes a tyrosine kinase receptor for vascular endothelial growth factors C and D. In humans, mutations in the gene cause Milroy disease (lymphedema) with the causative variants identified in patients with the disease occuring within the kinase domains (Connell *et al.*, 2009). However, Dumont et *al.* (1998) showed that the early inactivation of *Flt4* in mice leads to the abnormal development of large blood vessels, including the dorsal aorta. They concluded that “VEGFR-3 has an essential role in the development of the embryonic cardiovascular system before the emergence of the lymphatic vessels” (Dumont *et al.*, 1998). These results were corroborated by later studies in mice (Haiko *et al.*, 2008). Further mining the literature using IVA (QIAGEN, 2015) revealed that *FLT4* directly interacts with 3 genes that are all associated with cardiac defects found in ToF patients: *VEGF*, *NOS3* and *STAT3*. The VEGFR3 protein binds with the VEGFA protein (Kukk *et al.*, 1996). *Vegfa,* when knocked out in mice can lead to an overriding aorta, VSD and persistent truncus arteriosus (Stalmans *et al.*, 2003). In fact, Stalmans *et al.* (2003) explicitly describe cardiac defects in some of these mice as ToF. The VEGFR3 protein also increases activation of the eNOS protein (Lahdenranta *et al.*, 2009). *Nos3* knockout in mice has been known to lead to VSD (Feng *et al.*, 2002). Finally, the VEGFR3 protein increases activation of STAT3 protein (Korpelainen *et al.*, 1999). *Stat3* knockout in mice cardiac cells can lead to pulmonary stenosis (Zhang *et al.*, 2009). The different pathways from FLT4 to CHD are represented in Figure 28. *CCDC102B* by comparison, is not well characterized.

In addition to rare truncating SNVs having been found in *FLT4* in ToF patients, a single CNV encompassing the gene *FLT4* was also found in a study by Soemedi *et al*. (2012b) of 283 ToF trios; including probands analysed in this study. Soemedi *et al*. (2012b) found that 3 of the 5 genes covered by the duplication in 5q35.3 are expressed in fetal heart. The results of my study would suggest that it is the disruption of *FLT4* that leads to ToF in this patient.

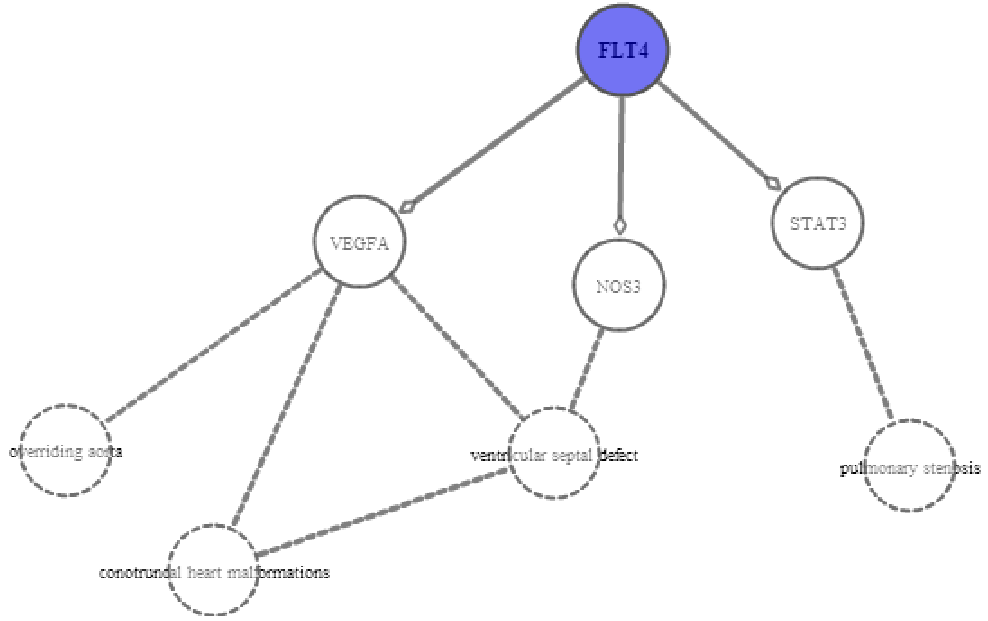

Figure 28. Pathway from FLT4 to CHDs relevant to ToF as seen in IVA (QIAGEN, 2015).

A large number of clusters were found for rare SNVs predicted deleterious. This included a cluster in *NOTCH1*, one of the genes in which we have previously established an overabundance of SNVs in ToF cases. Some of the top clustering results involved long genes with SNVs spread across many exons with no particular region showing a concentration of SNVs. This was the case for *HMCN1*, *MYO7A*, *NAV2* and *NEB*. None of these genes have been reported as involved in cardiogenesis. *HMCN1* mutations have so far been identified as the cause of age-related macular degeneration (Schultz *et al.*, 2003). *NEB* has been determined as the leading cause of nemaline myopathy (Lehtokari *et al.*, 2006). Mutations in *MYO7A* have led to cases of deafness and Usher syndrome; which leads to visual impairments in addition to deafness (NCBI, 2015). None of those genes appear to present links with cardiogenesis. On the other hand, 9 of 14 SNVs in *SERBP1* were found in the first exons, forming a cluster at both exon and gene level. Variants in *SERBP1* have not been linked to disease, but the protein product of the gene binds with the mRNA product of *SERPINE1*; a serine proteinase inhibitor, which has been implicated in thrombophilia; an increase in the risk of blood clotting (Rebhan *et al.*, 1998). The only indication of a large cardiac defect being associated with the mutated gene was in a patient presenting an inferior vena cava interruption; in addition to a deep vein thrombosis, who also had a variant in *SERPINE1* (Galati *et al.*, 2011). The patients also had several other variants in genes associated with thrombophilia (Galati *et al.*, 2011). *NELL1* mutations lead to congenital skeletal defects (Desai *et al.*, 2006). Cases of syndromic ToF have been accompanied by skeletal defects; as in Alagille syndrome

(Bauer *et al.*, 2010). However, no link between *NELL1* and congenital heart disease has been established yet. 6 SNVs were found in *SNCAIP* including 4 in a single exon. However, this exon is non-coding in the particular transcript for which clustering was detected (Cunningham *et al.*, 2015). *PTPN5* presents an interesting case as 6 SNVs were identified specifically in one of 15 exons. It has been suggested that another protein tyrosine phosphatase, *PTPN11,* influences the susceptibility of an individual to ToF; in this case through the action of a common variant (Goodship *et al.*, 2012).

One of the limitations of this study was the absence of a well-matched reference set which could act as a control, that is, individuals not selected for the presence of any specific condition, who had been sequenced using identical methodology. A study of the ToF patients which uses controls from the 1000 genomes project is currently underway. This study will follow some of the methodologies outlined in Section 5.2.2, which could not be applied to this current study. The results of these future studies will complement the study presented in this chapter. Further work will also involve the refinement of cluster detection strategies. The rate of variation differs for each gene and could have an impact on cluster detection. This should be taken into account for better cluster identification. The recent release of the repository ExAC offers one possible resource that could provide per gene estimates of variation rate.

5.6. Conclusion

At the time of writing, this study is the largest WES study of ToF yet reported. It involved the study of 824 ToF patients against 490 patients with various neurodevelopmental disorders. Despite the inherent limitations of the comparator set, I was able to uncover an overabundance of genes previously implicated in ToF in ToF patients compared to NDD patients. For one specific category, this extended to genes previous implicated in syndromes that involved ToF. However, these genes only accounted for a small fraction of ToF patients. I used clustering to uncover genes that may also be implicated in ToF. With support from pathway analysis and the existing literature on the genes found, I was able to uncover a promising candidate, *FLT4*. My study also strongly suggests that *NOTCH1* in involved in the manifestation of sporadic non-syndromic. This study also suggests that highly-penetrant variants can only explain a small fraction of ToF cases, leading to the conclusion that a more complex genetic etiology might be at work in many

ToF cases. Further refinements in methodology and in the choice of comparator set should provide more insight.

Chapter 6. General Discussion

6.2. Summary

The overarching aim of this PhD was to uncover potential causal rare variants in patients with congenital heart disease (CHD). My work had three aspects: the study of families in which a disease trait was segregating, the study of *de novo* mutation in sporadic cases of transposition of the great arteries (TGA) and the study of a large series of cases of tetralogy of Fallot (ToF). Each of these studies was conducted using WES and focused on variants found in protein-coding regions.

The study of CHD was initiated with the analysis of 8 families. 7 of these families exhibited cardiovascular disease, 5 with a CHD phenotype; the other two displaying relapsing cardiomyopathy and atypical Brugada syndrome (BrS) respectively. The 8 families had been analysed over the course of a few years and remained unresolved. There can be many possible reasons for the absence of suitable candidates in such projects, starting with the genuine absence of causal rare variants. For example, the disease could have been triggered by a rare copy number variant (CNV) or represented an aggregation of sporadic cases. Another reason could be the poor quality of sequencing at particular loci, leading the causal variant to be missed in one or more carriers. The design and implementation of BAMily was an attempt to address this second issue. I analysed all 8 families using BAMily and SAMtools. This was followed by the analysis of *de novo* mutation in sporadic cases of TGA. 32 probands were sequenced with their unaffected parent; and in one family an unaffected sibling. The cohort only included non-syndromic cases of TGA; cases that did not exhibit additional cardiac or extracardiac defects. The third analysis was based on a large series of 867 non-syndromic ToF cases. This study was originally envisioned as a case-control comparison. However, we were unable to obtain a control set sequenced at the same site as the non-syndromic ToF cases and with the same technology. I concentrated on finding clusters of variants within our cases and used as a reference set 500 cases from multiple studies of neurodevelopmental disorders (NDD) from the UK10K study (Wellcome Trust Sanger Institute, 2010). This was the closest dataset in terms of sequencing approach available to me at the time. I found the sets not to be suitably matched, but was nonetheless able to gain insights into the genetic etiology of ToF.

In this chapter, I review each of the three studies, including limitations and possible improvements, and integrate the main findings and discuss them in relation to the primary objective of my thesis.

6.3. Limitations.

6.3.1. Limitations working with WES data

In Chapter 3, I presented some of the limitations intrinsic to WES data. Specifically, I described how a new approach to variant calling could address issues arising from imperfect sequence coverage. I implemented and tested the variant caller BAMily which I subsequently applied to the 8 families in which a disease was segregating. These limitations also informed the design of the study of non-syndromic TGA. In this study, I obtained variants from the consensus of two variant callers, SAMtools (Li *et al.*, 2009a) and GATK UnifiedGenotyper (DePristo *et al.*, 2011) in order to limit the number of false positives. In each study, calling indels proved to be problematic. Many indels were picked out as false positives in IGV (Robinson *et al.*, 2011). I excluded indels from the core of the study of non-syndromic ToF patients. Indels were judged on a case by case basis in other studies.

6.3.2. Limitations due to study assumptions

Each WES analysis presented in this thesis was built on the assumption of single rare variants causing disease. The analysis of non-syndromic TGA and ToF patients was also based on the assumption that these variants were highly penetrant. Each assumption implies limitations on what results can be found.

In two families, no potential causal variants were identified. For these families, a more complex genetic etiology could be at work. Overall, likely causative SNVs and indels were found in 5 out of 8 families; or 62.5% of families. This rate of success is comparable to the 60% reported by Gilissen *et al*. (2012) for their WES projects centred on Mendelian disease. Extending the search for rare variants to CNVs led to the discovery of a causative CNV in one of the families.

The study of 32 TGA patients and their parents yielded 24 *de novo* variants across 14 probands, with 16 variants in 11 probands predicted as deleterious by at least one pathogenicity predictor. The rate of *de novo* variants per proband was 0.75, comparable to the estimated rate of *de novo* variants in the protein-coding exome on average per newborn of 0.74; determined from the rate for whole-genome (Veltman and Brunner, 2012b). If most TGA cases in this study arose through highly-deleterious *de novo* mutations, this would be reflected by the rate of *de novo* variants per proband due to effect of selection bias (Girard *et al.*, 2011). Instead, what the rate found suggests is that highly-deleterious *de novo* do not account for most TGA cases. This was the general conclusion put forward by Zaidi *et al*. (2013) for severe CHD. Specifically, Zaidi *et al*. (2013) estimated that *de novo* mutation could contribute 10% of severe CHD. In their study, Zaidi *et al*. (2013) compared the rate of *de novo* variants per individual in genes selected for their expression in the embryonic mouse heart cases and controls. With rates of 0.88 and 0.85 *de novo* variants per individual in cases and controls, Zaidi *et al*. (2013) determined that the difference between patients with severe CHD and controls was not statistically significant. The study also reported few *de novo* variants to which TGA could be attributed. For 65 TGA patients, no more than 3 potentially causative *de novo* mutations were found. In my own study I have concluded that *de novo* variants are not the major factor in the development of TGA. The current approach can be used to elucidate the few cases that are due to disease-causing *de novo* variants. Prospective disease-causative *de novo* variants that have been validated have been identified in 3 TGA patients; or around 10% of cases. However, a different approach will be needed to complement the genetic etiology of TGA.

The principal limiting factor in the study of non-syndromic ToF patients was the lack of a well-matched reference set, as discussed in Chapter 5. The rate of clustering in each cohort of patients nonetheless strongly suggested that my starting assumption did not apply to non-syndromic ToF either. The rates of clustering, when variants were only counted once per position, in truncating SNVs and SNVs predicted as deleterious were equivalent for ToF and NDD patients. This suggested one of two possibilities: either highly-deleterious rare SNVs influence the risk of disease to a high degree in both sets of patients or it does so in neither. There are at least two reasons to suspect the latter is true. One reason is that NDD have already been identified as polygenic disorders, arising from the combined action of several common and rare variants (Clarke *et al.*, 2015). Incidentally, this could explain the higher rate of clustering in NDD patients when each variant was counted for both truncating SNVs and SNVs predicted as deleterious with this approach favouring variants that are near the limit between common and rare variants. A second reason is the small role that SNVs in genes previously implicated in ToF play in this cohort. Genes previously implicated in ToF through gene sequencing studies showed no sign of clustering. Only 0.8% of ToF patients had truncating SNVs in these genes, 3.6% of ToF patients if variants predicted as deleterious were considered instead. In this latter case, only a fraction of the SNVs uncovered will truly be disease-causative. The relative absence of clusters in genes that can be associated with a CHD phenotype also suggests highly-deleterious rare SNVs cannot account for the majority of ToF cases. Clusters in *FLT4* and *NOTCH1* are the exception and as such, their finding is the main result of the study.

The assumptions made for cluster detection also place limits on this last study. It was assumed that the number of rare variants found in any segment of the exome of a set size would be the same regardless of locus. This leads to false clustering in genes that tolerate high mutation rates. In this study, I tackled this issue by excluding genes in which synonymous variants cluster. I also excluded genes with any type of clustering found in NDD patients. However, an approach that can avoid these false clusters in the first place is needed. I make suggestions of improvement in the following section.

6.3.3. Potential disease-causing variants.

The WES studies centred around cardiovascular disease have led to several outcomes. With the large series of non-syndromic ToF cases, the influence of genes implicated in ToF through gene-sequencing studies was reaffirmed, albeit with a smaller contribution to disease etiology than suggested by the original studies (see Table 1, Chapter 5). The role of mutation in *NOTCH1* in the development of ToF phenotypes was also reaffirmed (Wessels and Willems, 2010). While *SCN5A* and *SCN10A* have both been implicated in Brugada syndrome previously, the identification of a CNV spanning both genes constitutes a new finding (Hu *et al.*, 2014). For 7 genes in particular, evidence from the literature supports a role in cardiovascular disease: *CLTCL1*, *FLT4*, *GREB1*, *HES1*, *POPDC3*, *RBP5* and *SLC5A6*. There is some evidence to suggest that a rare *de novo* variant in *SNX13* in a patient with non-syndromic TGA could also be disease-causing although this is still a speculative conclusion. The discovery of disease-causing variants in families in which disease is segregating involved the design and implementation of a new variant caller: BAMily.

Variants in *NOTCH1* were found in multiple studies. A missense variant in *NOTCH1* was detected; and subsequently validated, in a family in which two siblings were diagnosed as having pulmonary atresia (PA) and ventricular septal defect (VSD) accompanied by major aortopulmonary collateral arteries (MAPCA). Interestingly, the combination of PA and VSD; occasionally with MAPCA, is typically interpreted as a variation of the ToF phenotype (Prieto, 2005). Additionally, 5 nonsense variants were found in *NOTCH1* among non-syndromic ToF patients that were not found in NDD patients. *NOTCH1* was also found to harbour a cluster of single nucleotide variants (SNV) predicted as deleterious in ToF patient. No cluster was found among synonymous variants or in any category of SNV for NDD patients. While it is not possible to ascertain which specific SNVs predicted as deleterious are disease-causative, the truncating SNVs are likely to be disease-causative. *NOTCH1* has been implicated in ToF, in a single individual in a family case study and a in a single CNV in a study of ToF patients (Garg *et al.*, 2005; Greenway *et al.*, 2009). My results demonstrate that disease-causative variants for *NOTCH1* are found in non-syndromic ToF patients as well. Interestingly, the variant in the family I studied is assumed to have reduced penetrance as neither parents present the disease phenotype (see Chapter 3, Figure 6). It is therefore possible that some *NOTCH1* variants could also have reduced penetrance in non-syndromic ToF patients.

Another strong gene candidate in the development of ToF is *FLT4*. 8 nonsense variants were found to cluster within *FLT4* in ToF patients. The protein product of *FLT4* interacts with the protein from 3 genes; *VEGF*, *NOS3* and STAT3, through binding and increased activation (Feng *et al.*, 2002; Stalmans *et al.*, 2003; Zhang *et al.*, 2009). These genes have been implicated in cardiac defects found in ToF. Also of interest is the fact that the expression of *STAT3* is also regulated by another gene previously implicated in ToF, *PTPN11* (Zhang *et al.*, 2009; Goodship *et al.*, 2012). A common variant in *PTPN11* was shown to influence the risk of developing ToF (Goodship *et al.*, 2012). *FLT4* and *PTPN11* could be influencing disease risk through *STAT3*.

A *de novo* missense variant in *RBP5* was found in a patient with non-syndromic TGA. A variant in that gene had previously been implicated in a case of total anomalous venous return (Nash *et al.*, 2015). Investigating this result, Nash *et al*. (2015) created Zebrafish mutants in *rbp7a*, the closest homolog to *RBP5*. In almost half of fish, the mutation led to abnormal looping of the heart (Nash *et al.*, 2015). The fact that the gene encodes a retinol-binding protein is also particularly significant given that the intake of retinol supplements during pregnancy or the injection of retinoic acid in a pregnant mouse triggers TGA in offspring (Loffredo *et al.*, 2001).

Two missense variants, in *HES1* and *PODC3*, were found in a pair of siblings with double-outlet right ventricle (DORV). The most promising candidate was the variant in *HES1*. Like *NOTCH1* and *JAG1*, *HES1* is a component of the Notch signalling pathway (Rochais *et al.*, 2009). The homolog *Hes1* is expressed in the second heart field during cardiogenesis in mice, playing an essential role in outflow tract development (Rochais *et al.*, 2009). *POPDC3* shares sequence similarity with *BVES*, a gene implicated in ToF through gene sequencing studies (Wu *et al.*, 2013). My own study of non-syndromic ToF cases does not support the notion that BVES plays a substantial role in ToF which casts some doubt on the role of *POPDC3* on DORV.

For two missense variants, in genes *CLTCL1* and *GREB1*, evidence of a potential role in cardiac malformation was obtained from CNV studies. A variant in *CLTCL1* was found in a mother and daughter presenting complex CHD profiles which included extracardiac features; such as lymphoedema in the mother. *CLTCL1* is one of the genes lost as a result of a 22q11.2 deletion (Michaelovsky *et al.*, 2012). Crucially, the daughter had an interrupted aortic arch (IAA). Half of IAA cases are attributable to a 22q11.2 deletion syndrome (Kobrynski and Sullivan, 2007). Although the deletion of *TBX1* has been typically designated as the main contributor of cardiac defects in 22q11.2 deletion syndrome, a role for *CLTCL1* cannot be excluded. A *de novo* missense variant in *GREB1* was found in a non-syndromic TGA patient. In a study conducted by Fakhro *et al*. (2011), *GREB1* was lost as a result of a rare CNV in a case of malposition of the great arteries (Fakhro *et al.*, 2011). Expression patterns in *Xenopus Tropicalis* suggested a role in cardiogenesis, specifically left-right patterning (Fakhro *et al.*, 2011).

One non-syndromic TGA patient has a *de novo* missense variant in *SNX13*. Knockouts of the mouse homolog *SNX13* lead to abnormal blood vessel morphology and embryonic lethality (Zheng *et al.*, 2006). There is potentially a link between sorting nexins and heterotaxy; a feature of syndromic TGA cases, via ciliopathies (Ware *et al.*, 2011; Chen *et al.*, 2012). However, this line of reasoning remains highly speculative at this stage.

Two of the three WES study designs used to uncover disease-causing variants made use of BAMily. The design, implementation and testing of the variant caller are covered at length in Chapter 3. In addition to laying the foundation for a new approach to variant calling, BAMily has contributed to the discovery of potentially disease-causative rare variants in families with a disease that is segregating and in non-syndromic TGA patients.

6.3.4. Outlook

The WES analysis of patients with various cardiac defects throughout this PhD has led to the discovery of likely disease-causing variants. In due course, these will be studied in biological experiments investigating the mechanisms that lead to heart disease. The necessity for a new approach to variant calling has also led to the design and implementation of the variant caller BAMily, which has contributed to variant discovery in two analyses. However, what these analyses have also revealed is that only a small fraction of non-syndromic TGA and ToF cases can be attributed to highly-penetrant rare variant. This is consistent with what the literature already suggests for severe CHD, whether SNVs or CNVs are considered (Greenway *et al.*, 2009; Soemedi *et al.*, 2012b; Zaidi *et al.,* 2013). Elucidating the genetic etiology of sporadic CHD will therefore require the study of rare variants that influence the risk of disease with varying effect sizes. Methods for collapsing rare variants might provide some insight in the short-term (Lee *et al.*, 2014b). However, detecting the effect of single rare variants that are not fully penetrant will require large case-controls studies with a number of patients vastly exceeding what has been done for CHD in terms of WES studies to date (Zuk *et al.*, 2014).

6.4. Conclusion

The overarching aim of this PhD was to discover likely causal rare variants in patients with CHD as well as a few other congenital defects. The work I carried out towards fulfilling this aim revealed promising candidates across 9 genes: *CLTCL1*, *HES1*, *POPDC3*, *NOTCH1*, *FLT4*, *SLC5A6*, *GREB1*, *RBP5* and *SNX13*. With the exception of *NOTCH1*, these genes have not previously been implicated in the diseases of the patients in which candidates were found. *NOTCH1* harbours likely disease-causing variants in two distinct studies, a familial case of PA/VSD/MAPCA and patients with sporadic non-syndromic ToF. Many of the genes in which likely causative variants were uncovered for this thesis will be the subject of biological experiments. In the process of identifying potential causal rare variants, I have established that highly-deleterious mutations do not contribute to the majority of non-syndromic ToF and TGA cases, suggesting a more complex genetic etiology. Future studies of CHD will undoubtedly focus on elucidating the genetic contribution of variants with reduced penetrance on severe CHD phenotypes such as ToF and TGA.

Appendix

Full results of variant caller comparisons

The following tables contain supplementary data relevant to Section 3.4.4.

	SAMtools	GATK	FamSeq	PolyMutt	BAMily
Variants present in cousin-uncle pair, absent in unrelated individual (3 trios)					
SNVs (avg.)	6885	7313	7792	7159	7386
Position in calls	1181	1186	1203	1183	1186
True positive	1168	1179	1182	1176	1179
Sensitivity %	96.88	97.79	97.98	97.51	97.73
Specificity %	99.9	99.95	99.83	99.94	99.94
False discovery %	1.05	0.53	1.8	0.62	0.62
…in cousin-cousin pair, absent in unrelated individual (3 trios)					
SNVs (avg.)	6051	6459	6954	6301	6533
Position in calls	1100	1104	1124	1101	1105
True positive	1087	1098	1100	1096	1098
Sensitivity %	96.74	97.72	97.86	97.51	97.72
Specificity %	99.9	99.95	99.81	99.96	99.95
False discovery %	1.12	0.54	2.14	0.45	0.6
…in uncle-cousins trio, absent in unrelated individual (3 quartets)					
SNVs (avg.)	4261	4484	5035	4432	4476
Position in calls	780	778	799	778	782
True positive	769	774	776	773	774
Sensitivity %	96.85	97.48	97.82	97.35	97.52
Specificity %	99.91	99.96	99.83	99.96	99.94
False discovery %	1.5	0.6	2.8	0.69	0.98
…in cousins trio, absent in unrelated individual (1 quartet)					
SNVs	3731	3950	4309	3901	3971
Position in calls	716	710	724	709	714
True positive	703	706	707	705	706
Sensitivity %	96.7	97.11	97.25	96.97	97.11
Specificity %	99.9	99.97	99.87	99.97	99.94
False discovery %	1.82	0.56	2.35	0.56	1.12
…in uncle-cousins quartet, absent in unrelated individual (1 quintet)					
SNVs	2989	3113	3515	3094	3609
Position in calls	561	556	570	556	558
True positive	550	552	554	551	552
Sensitivity %	96.66	97.01	97.36	96.84	97.01
Specificity %	99.92	99.97	99.88	99.96	99.96
False discovery %	1.96	0.72	2.81	0.9	1.08

Table S1. Sequence data and microarray data comparison for various sample sizes with cousin-cousin and uncle-cousins arrangements shown separately. In this table, positions refer to the positions covered by the microarray data.

	SAMtools	GATK	FamSeq	PolyMutt	BAMily
Variants in cousins, absent in uncle and unrelated individual (1 quintet)					
SNVs	752	833	828	818	829
Position in calls	156	154	154	154	155
True positive	152	152	152	152	152
Sensitivity %	96.2	96.2	96.2	96.2	96.2
Specificity %	99.97	99.99	99.99	99.99	99.98
False discovery %	2.56	1.3	1.3	1.3	1.94
…in uncle-cousins trio, absent in one cousin and unrelated (3 quintets)					
SNVs (avg.)	1278	1360	1549	1360	1354
Position in calls	219	221	229	221	223
True positive	216	219	220	219	220
Sensitivity %	95.99	97.18	97.92	97.18	97.77
Specificity %	99.98	99.99	99.94	99.99	99.98
False discovery %	1.37	0.91	3.64	0.91	1.33
…in cousin pair, absent in uncle-cousin-unrelated trio (3 quintets)					
SNVs (avg.)	1095	1152	1165	1153	1186
Position in calls	169	170	170	170	171
True positive	166	169	168	169	169
Sensitivity %	96.52	97.85	97.47	97.85	97.85
Specificity %	99.98	99.99	99.98	99.99	99.99
False discovery %	1.38	0.6	1.38	0.6	1.17
…in uncle-cousin pair, absent in cousins-unrelated trio (3 quintets)					
SNVs (avg.)	1410	1506	1586	1512	1529
Position in calls	183	184	188	184	184
True positive	180	183	183	183	182
Sensitivity %	95.38	97.03	97.38	97.18	96.84
Specificity %	99.98	99.99	99.96	99.99	99.99
False discovery %	1.83	0.5	2.65	0.69	0.69

Table S2. Sequence data and microarray data comparison for a fixed sample size of five, assuming different pairs and trios of individuals to be non-carriers. In this table, positions refer to the positions covered by microarray data.

Analysis of protein-protein functional links

The following table contain supplementary data relevant to Section 4.4.1.

	PHLPP2	PDE4D	ENTPD2	ZNF419	ZNF227	GPR17	RAD52	CREB3L3	ZFHX3	KCNJ12	FAM208B	CPT2	PROK1	GREB1	HECTD4	ANGPTL2	COL11A2	ZNF577	DCDC1	RBP5	SNX13	LMOD3	GDPGP1	KRTAP4-6
PHLPP2	■	1	2	2	1	2	2	2	2	2	2	3	2	2	2	2	2	2	3	3	3	3	3	4
PDE4D	1	■	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	2	2	3	4
ENTPD2	2	2	■	3	3	2	2	3	3	3	3	3	2	3	3	3	3	3	2	2	3	3	3	4
ZNF419	2	2	3	■	2	3	3	3	2	3	2	3	3	3	2	3	3	2	3	3	3	4	3	4
ZNF227	1	2	3	2	■	3	2	3	3	3	3	3	3	3	2	3	3	2	3	3	3	4	4	4
GPR17	2	2	2	3	3	■	3	3	3	2	3	3	1	3	3	3	3	3	3	3	3	3	4	4
RAD52	2	2	2	3	2	3	■	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5
CREB3L3	2	2	3	3	3	3	3	■	3	2	3	2	3	2	3	2	3	3	3	3	3	3	4	5
ZFHX3	2	2	3	2	3	3	2	3	■	3	2	3	3	3	3	3	3	3	3	3	3	4	3	4
KCNJ12	2	2	3	3	3	2	3	2	3	■	3	3	3	3	3	3	3	3	3	3	3	3	4	4
FAM208B	2	2	3	2	3	3	3	3	2	3	■	3	3	3	3	3	3	3	3	3	3	4	3	4
CPT2	3	2	3	3	3	3	3	2	3	3	3	■	3	3	3	2	2	3	3	3	3	4	3	5
PROK1	2	2	2	3	3	1	3	3	3	3	3	3	■	3	3	3	3	4	3	3	3	3	4	5
GREB1	2	2	3	3	3	3	3	2	3	3	3	3	3	■	3	3	3	3	3	3	3	3	4	5
HECTD4	2	3	3	2	2	3	3	3	3	3	3	3	3	3	■	3	3	2	3	3	3	4	4	5
ANGPTL2	2	3	3	3	3	3	3	2	3	3	3	2	3	3	3	■	3	4	3	3	3	3	4	5
COL11A2	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	■	3	3	3	4	3	4	4
ZNF577	2	3	3	2	2	3	3	3	3	3	3	3	4	3	2	4	3	■	3	4	3	4	4	5
DCDC1	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	■	3	4	4	4	5
RBP5	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	3	■	4	4	4	5
SNX13	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	3	4	4	■	4	4	5
LMOD3	3	2	3	4	4	3	3	3	4	3	4	4	3	3	4	3	3	4	4	4	4	■	5	5
GDPGP1	3	3	3	3	4	4	3	4	3	4	3	3	4	4	4	4	4	4	4	4	4	5	■	5
KRTAP4-6	4	4	4	4	4	4	5	5	4	4	4	5	5	5	5	5	4	5	5	5	5	5	5	■

Figure S1. Number of functional interactions separating protein products of genes from the final list of 24.

Tables that include both SNVs and indels.

The following tables contain supplementary data relevant to Section 5.4.2.

Gene set	Carrier frequency in ToF cohort (%)	Number of genes	Number of variants	Carrier frequency in NDD cases (%)	Number of genes	Number of variants
Known ToF	1.8	6	15	0	0	0
IVA	5	17	42	0.6	3	3
ToF syndrome	6.7	32	64	3.1	13	15
HRC	19	127	221	13.9	53	71
Zaidi	18	116	224	16.1	56	88
CHD syndrome	38.5	284	587	33.5	145	213
Cilia	10	36	104	6.3	21	31

Table S3. Rare truncating variants in ToF and NDD cases present in genes from CHD gene sets.

Gene sets	Carrier frequency in ToF cases (%)	Number of genes	Number of variants	Carrier frequency in NDD cases (%)	Number of genes	Number of variants
Known ToF	5.3	16	47	2.2	7	11
IVA	11	35	102	8.2	16	42
ToF syndrome	33.1	59	378	19.4	44	106
HRC	69.4	239	1113	61	178	474
Zaidi	64.6	194	984	56.7	149	420
CHD syndrome	92.5	490	2601	90	404	1191
Cilia	**31.9**	**45**	**334**	**33.9**	**43**	**208**

Table S4. Rare variants predicted to be deleterious in ToF and NDD cases present in genes from CHD gene sets.

Updated tables following additional filtering against 875 controls from the 1000g project

The following tables contain supplementary data relevant to Section 5.4.3.

CDS type	TOF cases (n=824)	NDD case (n=490)
Truncating	6,830	2,680
Gene	24	28
Exon	42	60
Both	18	17
Predicted deleterious	33,998	17,131
Gene	45	25
Exon	107	60
Both	19	17
Non-synonymous	71,541	34,844
Gene	62	21
Exon	96	49
Both	28	5
Synonymous	65,154	27,442
Gene	83	13
Exon	255	57
Both	**50**	**7**

Table S5. The number of genes with significant clustering (by variant) at gene and exon levels following additional filtering against 875 controls from the 1000g project. The total number of positions in each category in ToF and NDD cases is given in the header of each section. Only SNVs are considered here

Complete tables for Predicted Deleterious

Gene	**Clustering level**	**Cluster size (*p*-value)**
ADAMTS17*	Gene	17 (4.15E-3)
APC2*	Exon	5 (6.24E-3)
ATG2A	Gene	10 (1.17E-3)
CC2D2A	Exon	4 (4.07E-2)
CCDC101	Gene	8 (3.46E-2)
COL3A1	Gene	19 (1.62E-2)
CPXM2	Gene	14 (4.11E-3)
CSMD3	Gene	30 (7.33E-3)
FLNB	Gene/ Exon	29 (2.66E-4) / 5 (2.48E-2)
GLDC	Gene	20 (9.11E-6)
H6PD	Gene	15 (1.27E-3)
HMCN1	Gene	56 (4.37E-7)
HOOK1	Gene	13 (1.48E-2)
INSC	Gene	11 (1.40E-2)
KIF16B	Gene	8 (1.57E-3)
LARS	Gene	11 (6.82E-3)
MCM3AP	Exon	5 (4.76E-2)
MEF2D*	Gene	6 (2.00E-2)
MYO7A	Gene	30 (6.36E-6)
MYOM1	Exon	5 (3.19E-3)
NAV2	Gene	32 (6.38E-6)
NEB	Gene	58 (2.83E-5)
NECAB2	Gene	8 (4.54E-2)
NEK2	Gene	10 (2.09E-2)
NELL1	Gene	17 (4.80E-5)
NMNAT1	Gene/ Exon	8 (2.33E-2)/ 6 (5.90E-4)
NOTCH1	Gene	31 (7.89E-5)
PCDH9	Exon	4 (3.79E-4)
POLRMT	Exon	9 (1.64E-2)
PORCN	Exon	4 (4.72E-2)
PRRC2C	Gene	9 (4.36E-3)
PTPN5	Exon	6 (1.35E-4)
PUS7	Exon	4 (1.24E-2)
RANBP9	Exon	6 (1.01E-3)
RYR3	Gene	40 (2.65E-2)
SERBP1	Gene/ Exon	14 (6.18E-7)/ 9 (1.92E-5)
SLC26A1*	Gene	14 (1.59E-3)
SNCAIP	Gene	6 (8.95E-5)/ 4 (2.99E-2)
ST13	Exon	4 (4.95E-2)
STX6*	Gene	8 (1.12E-2)
TARBP1	Exon	4 (1.60E-2)
VCL*	Gene	13 (4.74E-2)
ZFR	Exon	5 (7.68E-3)
ZNF385D	Exon	5 (7.07E-3)
ZNF423	Exon	8 (6.61E-10)

Table S6. Size and p-value (by position) of SNVs predicted to be deleterious in ToF patients. Underlined are genes that are also present in NDD cases or the Synonymous category. *Gene no longer in list after stringent filtering against 875 controls from the 1000g project.

Tables of genes with clusters and in gene sets.

The following tables contain supplementary data relevant to Section 5.4.4.

Gene sets	Predicted Deleterious	Non-Synonymous	Synonymous	In NDD cases
ToF syndrome	*NOTCH1*, *PORCN*			
HRC	*MEF2D**, *MYOM1*, *NOTCH1*, *VCL**	*LAMA5*, *MESP1**	*LAMA5*	
Zaidi	*NEK2*, *NOTCH1*	*PCNT*		*VIT*
CHD syndrome	*ADAMTS17**, *CC2D2A*, *COL3A1*, *NOTCH1*, *PORCN*, *ZNF423*	*HSPG2**, *PCNT*, *PLEC*, *ZNF423*	*MECP2*, *PLEC*, *RYR1*, *ZNF423*	*CENPJ*, *COL17A1**, *HSPG2*, *PLEC**, *ZNF423*
Cilia	*CC2D2A*			

Table S7. genes from CHD gene sets that harbour clusters (by position) for ToF cases. In bold are genes appearing in multiple gene sets (except shared between ToF and CHD syndrome sets as one includes the other). Underlined are genes that are also present in NDD cases or the Synonymous category. *Gene no longer in list if stringent filtering is applied.

References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), pp. 56-65.

Adler, L.N., Gomez, T.A., Clarke, S.G. and Linster, C.L. (2011) 'A novel GDP-D-glucose phosphorylase involved in quality control of the nucleoside diphosphate sugar pool in Caenorhabditis elegans and mammals', *Journal of Biological Chemistry*, 286(24), pp. 21511-23.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) 'A method and server for predicting damaging missense mutations', *Nat Meth*, 7(4), pp. 248-249.

Agilent Technologies (2015) *SureSelect All Exon Kits Details & Specifications.* Available at: http://www.genomics.agilent.com/article.jsp?pageId=3042 (Accessed: 8th of August).

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biology*, 12(2).

Alizadehasl, A. and Sadeghpour, A. (2014) 'Congenital Aortic Valve Stenosis', in *Comprehensive Approach to Adult Congenital Heart Disease*. Springer, pp. 275-279.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. and Müller-Myhsok, B. (2012) 'A beginners guide to SNP calling from high-throughput DNA-sequencing data', *Human Genetics*, 131(10), pp. 1541-1554.

Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E.B. and Muller-Myhsok, B. (2011) 'vipR: variant identification in pooled DNA using R', *Bioinformatics*, 27(13), pp. I77-I84.

Altshuler, D., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De la Vega, F.M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S.B., Gibbs, R.A., Knoppers, B.M., Lander, E.S., Lehrach, H., Mardis, E.R., McVean, G.A., Nickerson, D., Peltonen, L., Schafer, A.J., Sherry, S.T., Wang, J., Wilson, R.K., Gibbs, R.A., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J.X., Jian, M., Li, G., Li, R.Q., Liang, H.Q., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H.M., Zhang, X.Q., Zheng, H.S., Lander, E.S., Altshuler, D.L., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T.J., Gabriel, S.B., Jaffe, D.B., Shefler, E., Sougnez, C.L., Bentley, D.R., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K.J., Costa, G.L., Ichikawa, J.K., Lee, C.C., Sudbrak, R., Lehrach, H., Borodina, T.A., Dahl, A., Davydov, A.N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A.V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., et al. (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061-1073.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Gibbs, R.A., Dinh, H., Kovar, C., Lee, S., Lewis, L., Muzny, D., Reid, J., Wang, M., Wang, J., Fang, X.D., Guo, X.S., Jian, M., Jiang, H., Jin, X., Li, G.Q., Li, J.X., Li, Y.R., Li, Z., Liu, X., Lu, Y., Ma, X.D., Su, Z., Tai, S.S., Tang, M.F., Wang, B., Wang, G.B., Wu, H.L., Wu, R.H., Yin, Y., Zhang, W.W., Zhao, J., Zhao, M.R., Zheng, X.L., Zhou, Y., Lander, E.S., Altshuler, D.M., Gabriel, S.B., Gupta, N., Flicek, P., Clarke, L., Leinonen, R., Smith, R.E., Zheng-Bradley, X., Bentley, D.R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.L., Sherry, S.T., McVean, G.A., Mardis, E.R., Wilson, R.K., Fulton, L., Fulton, R., Weinstock, G.M., Durbin, R.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., et al. (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), pp. 56-65.

Andersen, T.A., Troelsen Kde, L. and Larsen, L.A. (2014) 'Of mice and men: molecular genetics of congenital heart disease', *Cell Mol Life Sci*, 71(8), pp. 1327-52.

Anderson, R.H. and Weinberg, P.M. (2005) 'the clinical anatomy of tetralogy of fallot', *Cardiology in the Young*, 15(s1), pp. 38-47.

Andree, B., Hillemann, T., Kessler-Icekson, G., Schmitt-John, T., Jockusch, H., Arnold, H.H. and Brand, T. (2000) 'Isolation and characterization of the novel Popeye gene family expressed in skeletal muscle and heart', *Developmental Biology*, 223(2), pp. 371-382.

Antonarakis, S.E. and Beckmann, J.S. (2006) 'Mendelian disorders deserve more attention', *Nat Rev Genet*, 7(4), pp. 277-82.

Antzelevitch, C., Brugada, P., Borggrefe, M., Brugada, J., Brugada, R., Corrado, D., Gussak, I., LeMarec, H., Nademanee, K., Perez Riera, A.R., Shimizu, W., Schulze-Bahr, E., Tan, H. and Wilde, A. (2005) 'Brugada syndrome: report of the second consensus conference: endorsed by the Heart Rhythm Society and the European Heart Rhythm Association', *Circulation*, 111(5), pp. 659-70.

Apitz, C., Webb, G.D. and Redington, A.N. (2009) 'Tetralogy of Fallot', *Lancet*, 374(9699), pp. 1462-71.

Appel, S., Filter, M., Reis, A., Hennies, H.C., Bergheim, A., Ogilvie, E., Arndt, S., Simmons, A., Lovett, M., Hide, W., Ramsay, M., Reichwald, K., Zimmermann, W. and Rosenthal, A. (2002) 'Physical and transcriptional map of the critical region for keratolytic winter erythema (KWE) on chromosome 8p22-p23 between D8S550 and D8S1759', *European Journal of Human Genetics*, 10(1), pp. 17-25.

Auer, P.L. and Lettre, G. (2015) 'Rare variant association studies: considerations, challenges and opportunities', *Genome Medicine*, 7.

Baban, A., Postma, A.V., Marini, M., Trocchio, G., Santilli, A., Pelegrini, M., Sirleto, P., Lerone, M., Albanese, S.B., Barnett, P., Boogerd, C.J., Dallapiccola, B., Digilio, M.C., Ravazzolo, R. and Pongiglione, G. (2014) 'Identification of TBX5 mutations in a series of 94 patients with Tetralogy of Fallot', *Am J Med Genet A*, 164A(12), pp. 3100-7.

Bailey-Wilson, J.E. and Wilson, A.F. (2011) 'Linkage analysis in the next-generation sequencing era', *Hum Hered*, 72(4), pp. 228-36.

Bailliard, F. and Anderson, R. (2009) 'Tetralogy of Fallot', *Orphanet Journal of Rare Diseases*, 4(1), p. 2.

Bajolle, F., Zaffran, S., Kelly, R.G., Hadchouel, J., Bonnet, D., Brown, N.A. and Buckingham, M.E. (2006) 'Rotation of the Myocardial Wall of the Outflow Tract Is Implicated in the Normal Positioning of the Great Arteries', *Circulation Research*, 98(3), pp. 421-428.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. (2011) 'Exome sequencing as a tool for Mendelian disease gene discovery', *Nat Rev Genet*, 12(11), pp. 745-755.

Bansal, V. (2010) 'A statistical method for the detection of variants from next-generation resequencing of DNA pools', *Bioinformatics*, 26, pp. i318 - i324.

Bansal, V., Libiger, O., Torkamani, A. and Schork, N.J. (2010) 'Statistical analysis strategies for association studies involving rare variants', *Nat Rev Genet*, 11(11), pp. 773-785.

Basson, C.T., Bachinsky, D.R., Lin, R.C., Levi, T., Elkins, J.A., Soults, J., Grayzel, D., Kroumpouzou, E., Traill, T.A., Leblanc-Straceski, J., Renault, B., Kucherlapati, R., Seidman, J.G. and Seidman, C.E. (1997) 'Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome', *Nat Genet*, 15(1), pp. 30-5.

Bauer, R.C., Laney, A.O., Smith, R., Gerfen, J., Morrissette, J.J.D., Woyciechowski, S., Garbarini, J., Loomes, K.M., Krantz, I.D., Urban, Z., Gelb, B.D., Goldmuntz, E. and Spinner, N.B. (2010) 'Jagged1 (JAG1) mutations in patients with tetralogy of fallot or pulmonic stenosis', *Human Mutation*, 31(5), pp. 594-601.

Benjamin, E.J., Rice, K.M., Arking, D.E., Pfeufer, A., van Noord, C., Smith, A.V., Schnabel, R.B., Bis, J.C., Boerwinkle, E., Sinner, M.F., Dehghan, A., Lubitz, S.A., D'Agostino, R.B., Lumley, T., Ehret, G.B., Heeringa, J., Aspelund, T., Newton-Cheh, C., Larson, M.G., Marciante, K.D., Soliman, E.Z., Rivadeneira, F., Wang, T.J., Eiriksdottir, G., Levy, D., Psaty, B.M., Li, M., Chamberlain, A.M., Hofman, A., Vasan, R.S., Harris, T.B., Rotter, J.I., Kao, W.H.L., Agarwal, S.K., Stricker, B.H.C., Wang, K., Launer, L.J., Smith, N.L., Chakravarti, A., Uitterlinden, A.G., Wolf, P.A., Sotoodehnia, N., Kottgen, A., van Duijn, C.M., Meitinger, T., Mueller, M., Perz, S., Steinbeck, G., Wichmann, H.E., Lunetta, K.L., Heckbert, S.R., Gudnason, V., Alonso, A., Kaab, S., Ellinor, P.T. and Witteman, J.C.M. (2009) 'Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry', *Nature Genetics*, 41(8), pp. 879-881.

Bernasconi, A., Azancot, A., Simpson, J.M., Jones, A. and Sharland, G.K. (2005) 'Fetal dextrocardia: diagnosis and outcome in two tertiary centres', *Heart*, 91(12), pp. 1590-1594.

Bernier, P.-L., Stefanescu, A., Samoukovic, G. and Tchervenkov, C.I. (2010) 'The Challenge of Congenital Heart Disease Worldwide: Epidemiologic and Demographic Facts', *Seminars in Thoracic and Cardiovascular Surgery: Pediatric Cardiac Surgery Annual*, 13(1), pp. 26-34.

Bezzina, C.R., Barc, J., Mizusawa, Y., Remme, C.A., Gourraud, J.-B., Simonet, F., Verkerk, A.O., Schwartz, P.J., Crotti, L., Dagradi, F., Guicheney, P., Fressart, V., Leenhardt, A., Antzelevitch, C., Bartkowiak, S., Borggrefe, M., Schimpf, R., Schulze-Bahr, E., Zumhagen, S., Behr, E.R., Bastiaenen, R., Tfelt-Hansen, J., Olesen, M.S., Kaab, S., Beckmann, B.M., Weeke, P., Watanabe, H., Endo, N., Minamino, T., Horie, M., Ohno, S., Hasegawa, K., Makita, N., Nogami, A., Shimizu, W., Aiba, T., Froguel, P., Balkau, B., Lantieri, O., Torchio, M., Wiese, C., Weber, D., Wolswinkel, R., Coronel, R., Boukens, B.J., Bezieau, S., Charpentier, E., Chatel, S., Despres, A., Gros, F., Kyndt, F., Lecointe, S., Lindenbaum, P., Portero, V., Violleau, J., Gessler, M., Tan, H.L., Roden, D.M., Christoffels, V.M., Le Marec, H., Wilde, A.A., Probst, V., Schott, J.-J., Dina, C. and Redon, R. (2013) 'Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death', *Nat Genet*, 45(9), pp. 1044-1049.

Biesecker, L.G. and Spinner, N.B. (2013) 'A genomic view of mosaicism and human disease', *Nat Rev Genet*, 14(5), pp. 307-320.

Bittel, D.C., Zhou, X.G., Kibiryeva, N., Fiedler, S. and O'Brien, J.E. (2014) 'Ultra High-Resolution Gene Centric Genomic Structural Analysis of a Non-Syndromic Congenital Heart Defect, Tetralogy of Fallot (vol 9, e87472, 2014)', *PLoS One*, 9(5).

Blue, G.M., Kirk, E.P., Sholler, G.F., Harvey, R.P. and Winlaw, D.S. (2012) 'Congenital heart disease: current knowledge about causes and inheritance', *The Medical Journal of Australia*, 197(3), pp. 155-159.

Bose, C.L. and Laughon, M.M. (2007) 'Patent ductus arteriosus: lack of evidence for common treatments', *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 92(6), pp. F498-F502.

Boshoff, D. and Gewillig, M. (2006) 'A review of the options for treatment of major aortopulmonary collateral arteries in the setting of tetralogy of Fallot with pulmonary atresia', *Cardiology in the Young*, 16(3), pp. 212-220.

Brognard, J., Sierecki, E., Gao, T. and Newton, A.C. (2007) 'PHLPP and a second isoform, PHLPP2, differentially attenuate the amplitude of Akt signaling by regulating distinct Akt isoforms', *Mol Cell*, 25(6), pp. 917-31.

Brugada, R., Campuzano, O., Brugada, P., Brugada, J. and Hong, K. (1993) *Brugada Syndrome* (2010/03/20). Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301690.

Bruneau, B.G. (2008) 'The developmental genetics of congenital heart disease', *Nature*, 451(7181), pp. 943-8.

Bruneau, B.G. and Srivastava, D. (2014) 'Congenital Heart Disease Entering a New Era of Human Genetics', *Circulation Research*, 114(4), pp. 598-599.

Brunham, L.R. and Hayden, M.R. (2013) 'Hunting human disease genes: lessons from the past, challenges for the future', *Human Genetics*, 132(6), pp. 603-17.

Buermans, H.P.J. and den Dunnen, J.T. (2014) 'Next generation sequencing technology: Advances and applications', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), pp. 1932-1941.

Burn, J., Brennan, P., Little, J., Holloway, S., Coffey, R., Somerville, J., Dennis, N.R., Allan, L., Arnold, R., Deanfield, J.E., Godman, M., Houston, A., Keeton, B., Oakley, C., Scott, O., Silove, E., Wilkinson, J., Pembrey, M. and Hunter, A.S. (1998) 'Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study', *The Lancet*, 351(9099), pp. 311-316.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., Todd, J.A., Donnelly, P., Barrett, J.C., Davison, D., Easton, D., Evans, D., Leung, H.T., Marchini, J.L., Morris, A.P., Spencer, C.C.A., Tobin, M.D., Attwood, A.P., Boorman, J.P., Cant, B., Everson, U., Hussey, J.M., Jolley, J.D., Knight, A.S., Koch, K., Meech, E., Nutland, S., Prowse, C.V., Stevens, H.E., Taylor, N.C., Walters, G.R., Walker, N.M., Watkins, N.A., Winzer, T., Jones, R.W., McArdle, W.L., Ring, S.M., Strachan, D.P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E.K., Grozeva, D., Hamshere, M.L., Holmans, P.A., Jones, I.R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M.C., Owen, M.J., Collier, D.A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A.H., Ferrier, I.N., Ball, S.G., Balmforth, A.J., Barrett, J.H., Bishop, D.T., Iles, M.M., Maqbool, A., Yuldasheva, N., Hall, A.S., Braund, P.S., Dixon, R.J., Mangino, M., Stevens, S., Thompson, J.R., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C.W., Nimmo, E.R., Satsangi, J., Fisher, S.A., Forbes, A., Lewis, C.M., Onnie, C.M., Prescott, N.J., Sanderson, J., Mathew, C.G., Barbour, J., Mohiuddin, M.K., Todhunter, C.E., Mansfield, J.C., Ahmad, T., Cummings, F.R., Jewell, D.P., et al. (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661-678.

Cemal, Y., Pusic, A. and Mehrara, B.J. (2011) 'Preventative measures for lymphedema: Separating fact from fiction', *Journal of the American College of Surgeons*, 213(4), pp. 543-551.

Chadwick, B.P. and Frischauf, A.M. (1997) 'Cloning and mapping of a human and mouse gene with homology to ecto-ATPase genes', *Mammalian Genome*, 8(9), pp. 668-672.

Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A. and Yu, F. (2012) 'An integrative variant analysis suite for whole exome next-generation sequencing data', *BMC Bioinformatics*, 13, p. 8.

Chang, X. and Wang, K. (2012) 'wANNOVAR: annotating genetic variants for personal genomes via the web', *Journal of Medical Genetics*, 49(7), pp. 433-436.

Chen, Q. and Sun, F. (2013) 'A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms', *BMC Genomics*, 14(Suppl 1), p. S1.

Chen, Y.Q., Wu, B., Xu, L.L., Li, H.P., Xia, J.H., Yin, W.G., Li, Z., Shi, D.W., Li, S., Lin, S., Shu, X.D. and Pei, D.Q. (2012) 'A SNX10/V-ATPase pathway regulates ciliogenesis in vitro and in vivo', *Cell Research*, 22(2), pp. 333-345.

Chilamakuri, C.S., Lorenz, S., Madoui, M.-A., Vodak, D., Sun, J., Hovig, E., Myklebost, O. and Meza-Zepeda, L. (2014) 'Performance comparison of four exome capture systems for deep sequencing', *BMC Genomics*, 15(1), p. 449.

Chun, S. and Fay, J.C. (2009) 'Identification of deleterious mutations within three human genomes', *Genome Research*, 19(9), pp. 1553-1561.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S. and Getz, G. (2013) 'Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples', *Nature Biotechnology*, 31(3), pp. 213-219.

Cipollone, D., Amati, F., Carsetti, R., Placidi, S., Biancolella, M., D'Amati, G., Novelli, G., Siracusa, G. and Marino, B. (2006) 'A multiple retinoic acid antagonist induces conotruncal anomalies, including transposition of the great arteries, in mice', *Cardiovascular Pathology*, 15(4), pp. 194-202.

Cirulli, E.T. and Goldstein, D.B. (2010) 'Uncovering the roles of rare variants in common disease through whole-genome sequencing', *Nat Rev Genet*, 11(6), pp. 415-425.

Clarke, T.K., Lupton, M.K., Fernandez-Pujals, A.M., Starr, J., Davies, G., Cox, S., Pattie, A., Liewald, D.C., Hall, L.S., MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Thomson, P.A., Hayward, C., Hansell, N.K., Montgomery, G.W., Medland, S.E., Martin, N.G., Wright, M.J., Porteous, D.J., Deary, I.J. and McIntosh, A.M. (2015) 'Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population', *Mol Psychiatry*.

Clyman, R.I. and Chorne, N. (2007) 'Patent ductus arteriosus: evidence for and against treatment', *The Journal of pediatrics*, 150(3), p. 216.

Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Res*, 38(6), pp. 1767-71.

Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. and Conso, I.H.G.S. (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931-945.

Connell, F.C., Ostergaard, P., Carver, C., Brice, G., Williams, N., Mansour, S., Mortimer, P.S., Jeffery, S. and Consortium, L. (2009) 'Analysis of the coding regions of VEGFR3 and VEGFC in Milroy disease and other primary lymphoedemas (vol 124, pg 625, 2009)', *Human Genetics*, 125(2), pp. 237-237.

Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., Zilversmit, M., Cartwright, R., Rouleau, G.A., Daly, M., Stone, E.A., Hurles, M.E. and Awadalla, P. (2011) 'Variation in genome-wide mutation rates within and between human families', *Nat Genet*, 43(7), pp. 712-4.

Cordell, H.J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J., Setchfield, K., Thornborough, C., Breckpot, J., Soemedi, R., Martin, R., Rahman, T.J., Hall, D., van Engelen, K., Moorman, A.F.M., Zwinderman, A.H., Barnett, P., Koopmann, T.T., Adriaens, M.E., Varro, A., George, A.L., dos Remedios, C., Bishopric, N.H., Bezzina, C.R., O'Sullivan, J., Gewillig, M., Bu'Lock, F.A., Winlaw, D., Bhattacharya, S., Devriendt, K., Brook, J.D., Mulder, B.J.M., Mital, S., Postma, A.V., Lathrop, G.M., Farrall, M., Goodship, J.A. and Keavney, B.D. (2013a) 'Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16', *Nat Genet*, 45(7), pp. 822-824.

Cordell, H.J., Topf, A., Mamasoula, C., Postma, A.V., Bentham, J., Zelenika, D., Heath, S., Blue, G., Cosgrove, C., Granados Riveron, J., Darlay, R., Soemedi, R., Wilson, I.J., Ayers, K.L., Rahman, T.J., Hall, D., Mulder, B.J., Zwinderman, A.H., van Engelen, K., Brook, J.D., Setchfield, K., Bu'Lock, F.A., Thornborough, C., O'Sullivan, J., Stuart, A.G., Parsons, J., Bhattacharya, S., Winlaw, D., Mital, S., Gewillig, M., Breckpot, J., Devriendt, K., Moorman, A.F., Rauch, A., Lathrop, G.M., Keavney, B.D. and Goodship, J.A. (2013b) 'Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot', *Hum Mol Genet*, 22(7), pp. 1473-81.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and D'Eustachio, P. (2014) 'The Reactome pathway knowledgebase', *Nucleic Acids Research*, 42(D1), pp. D472-D477.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R. and Flicek, P. (2015) 'Ensembl 2015', *Nucleic Acids Research*, 43(D1), pp. D662-D669.

D/'Alessandro, L.C.A., Al Turki, S., Manickaraj, A.K., Manase, D., Mulder, B.J.M., Bergin, L., Rosenberg, H.C., Mondal, T., Gordon, E., Lougheed, J., Smythe, J., Devriendt, K., Bhattacharya, S., Watkins, H., Bentham, J., Bowdin, S., Hurles, M.E. and Mital, S. (2015) 'Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect', *Genet Med*.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. and Grp, G.P.A. (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), pp. 2156-2158.

de la Cruz, M.V., Sanchez Gomez, C., Arteaga, M.M. and Arguello, C. (1977) 'Experimental study of the development of the truncus and the conus in the chick embryo', *J Anat*, 123(Pt 3), pp. 661-86.

De Luca, A., Sarkozy, A., Consoli, F., Ferese, R., Guida, V., Dentici, M.L., Mingarelli, R., Bellacchio, E., Tuo, G., Limongelli, G., Digilio, M.C., Marino, B. and Dallapiccola, B. (2010) 'Familial transposition of the great arteries caused by multiple mutations in laterality genes', *Heart*, 96(9), pp. 673-7.

Deanfield, J., Thaulow, E., Warnes, C., Webb, G., Kolbel, F., Hoffman, A., Sorenson, K., Kaemmerer, H., Thilen, U., Bink-Boelkens, M., Iserin, L., Daliento, L., Silove, E., Redington, A., Vouhe, P. and Cardiology, E.S. (2003) 'Management of grown up congenital heart disease', *European Heart Journal*, 24(11), pp. 1035-1084.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. and Daly, M.J. (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43(5), pp. 491-+.

Dering, C., Hemmelmann, C., Pugh, E. and Ziegler, A. (2011) 'Statistical analysis of rare sequence variants: an overview of collapsing methods', *Genetic Epidemiology*, 35(S1), pp. S12-S17.

Derkach, A., Chiang, T., Gong, J., Addis, L., Dobbins, S., Tomlinson, I., Houlston, R., Pal, D.K. and Strug, L.J. (2014) 'Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic', *Bioinformatics*, 30(15), pp. 2179-2188.

Desai, J., Shannon, M.E., Johnson, M.D., Ruff, D.W., Hughes, L.A., Kerley, M.K., Carpenter, D.A., Johnson, D.K., Rinchik, E.M. and Culiat, C.T. (2006) 'Nell1-deficient mice have reduced expression of extracellular matrix proteins causing cranial and vertebral defects', *Human Molecular Genetics*, 15(8), pp. 1329-1341.

Digilio, M.C., Casey, B., Toscano, A., Calabrò, R., Pacileo, G., Marasini, M., Banaudi, E., Giannotti, A., Dallapiccola, B. and Marino, B. (2001) 'Complete Transposition of the Great Arteries: Patterns of Congenital Heart Disease in Familial Precurrence', *Circulation*, 104(23), pp. 2809-2814.

Digilio, M.C., Marino, B., Banaudi, E., Marasini, M. and Dallapiccola, B. (1998) 'Familial recurrence of transposition of the great arteries', *The Lancet*, 351(9116), p. 1661.

Dong, C.L., Wei, P., Jian, X.Q., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X.M. (2015) 'Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies', *Human Molecular Genetics*, 24(8), pp. 2125-2137.

Dumont, D.J., Jussila, L., Taipale, J., Lymboussaki, A., Mustonen, T., Pajusola, K., Breitman, M. and Alitalo, K. (1998) 'Cardiovascular failure in mouse embryos deficient in VEGF receptor-3', *Science*, 282(5390), pp. 946-949.

Eldadah, Z.A., Hamosh, A., Biery, N.J., Montgomery, R.A., Duke, M., Elkins, R. and Dietz, H.C. (2001) 'Familial Tetralogy of Fallot caused by mutation in the jagged1 gene', *Human Molecular Genetics*, 10(2), pp. 163-169.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E. and Grp, M.G.D. (2015) 'The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease', *Nucleic Acids Research*, 43(D1), pp. D726-D736.

Ewert, P., Bertram, H., Breuer, J., Dähnert, I., Dittrich, S., Eicken, A., Emmel, M., Fischer, G., Gitter, R., Gorenflo, M., Haas, N., Kitzmüller, E., Koch, A., Kretschmar, O., Lindinger, A., Michel-Behnke, I., Nuernberg, J.H., Peuster, M., Walter, K., Zartner, P. and Uhlemann, F. (2011) 'Balloon valvuloplasty in the treatment of congenital aortic valve stenosis — A retrospective multicenter survey of more than 1000 patients', *International Journal of Cardiology*, 149(2), pp. 182-185.

Ewing, B. and Green, P. (1998) 'Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities', *Genome Research*, 8(3), pp. 186-194.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) 'Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment', *Genome Research*, 8(3), pp. 175-185.

Fahed, A.C., Gelb, B.D., Seidman, J.G. and Seidman, C.E. (2013) 'Genetics of congenital heart disease: the glass half empty', *Circulation Research*, 112(4), pp. 707-20.

Fakhro, K.A., Choi, M., Ware, S.M., Belmont, J.W., Towbin, J.A., Lifton, R.P., Khokha, M.K. and Brueckner, M. (2011) 'Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning', *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), pp. 2915-2920.

Feng, Q., Song, W., Lu, X., Hamilton, J.A., Lei, M., Peng, T. and Yee, S.P. (2002) 'Development of heart failure and congenital septal defects in mice lacking endothelial nitric oxide synthase', *Circulation*, 106(7), pp. 873-9.

Ferencz, C., Boughman, J.A., Neill, C.A., Brenner, J.I. and Perry, L.W. (1989) 'Congenital Cardiovascular Malformations - Questions on Inheritance', *Journal of the American College of Cardiology*, 14(3), pp. 756-763.

Fermilab (2014) *Scientific Linux*. Available at: https://www.scientificlinux.org/ (Accessed: 2nd of October).

Fesslova, V., Brankovic, J., Lalatta, F., Villa, L., Meli, V., Piazza, L. and Ricci, C. (2011) 'Recurrence of congenital heart disease in cases with familial risk screened prenatally by echocardiography', *J Pregnancy*, 2011, p. 368067.

Folli, C., Calderone, V., Ottonello, S., Bolchi, A., Zanotti, G., Stoppini, M. and Berni, R. (2001) 'Identification, retinoid binding, and x-ray analysis of a human retinol-binding protein', *Proc Natl Acad Sci U S A*, 98(7), pp. 3710-5.

Galati, G., Gentilucci, U.V., Mazzarelli, C., Gallo, P., Grasso, R.F., Stellato, L., Afeltra, A. and Picardi, A. (2011) 'Deep Vein Thrombosis, Inferior Vena Cava Interruption and Multiple Thrombophilic Gene Mutations', *American Journal of the Medical Sciences*, 342(1), pp. 79-82.

Gamss, C. and Haramati, L.B. (2014) 'Eisenmenger Syndrome', *Cardiac Imaging*, p. 154.

Garg, V., Kathiriya, I.S., Barnes, R., Schluterman, M.K., King, I.N., Butler, C.A., Rothrock, C.R., Eapen, R.S., Hirayama-Yamada, K., Joo, K., Matsuoka, R., Cohen, J.C. and Srivastava, D. (2003) 'GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5', *Nature*, 424(6947), pp. 443-7.

Garg, V., Muth, A.N., Ransom, J.F., Schluterman, M.K., Barnes, R., King, I.N., Grossfeld, P.D. and Srivastava, D. (2005) 'Mutations in NOTCH1 cause aortic valve disease', *Nature*, 437(7056), pp. 270-274.

Gelb, B.D. and Chung, W.K. (2014) 'Complex Genetics and the Etiology of Human Congenital Heart Disease', *Cold Spring Harb Perspect Med*, 4(7).

Gene Ontology, C. (2004) 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Research*, 32(suppl 1), pp. D258-D261.

Geva, T., Martins, J.D. and Wald, R.M. (2014) 'Atrial septal defects', *Lancet*, 383(9932), pp. 1921-32.

Ghosh, M.G., Thompson, D.A. and Weigel, R.J. (2000) 'PDZK1 and GREB1 are estrogen-regulated genes expressed in hormone-responsive breast cancer', *Cancer Res*, 60(22), pp. 6367-75.

Gilissen, C., Hoischen, A., Brunner, H.G. and Veltman, J.A. (2012) 'Disease gene identification strategies for exome sequencing', *European Journal of Human Genetics*, 20(5), pp. 490-497.

Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., Thibodeau, P., Bachand, I., Bao, J.Y.J., Tong, A.H.Y., Lin, C.-H., Millet, B., Jaafari, N., Joober, R., Dion, P.A., Lok, S., Krebs, M.-O. and Rouleau, G.A. (2011) 'Increased exonic de novo mutation rate in individuals with schizophrenia', *Nat Genet*, 43(9), pp. 860-863.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. and Valencia, A. (2012) 'EnrichNet: network-based gene set enrichment analysis', *Bioinformatics*, 28(18), pp. I451-I457.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S. and Nusbaum, C. (2009) 'Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing', *Nat Biotech*, 27(2), pp. 182-189.

GNU Project (2014) *Bash.* Available at: https://www.gnu.org/software/bash/ (Accessed: 8th of August).

Goldmuntz, E., Bamford, R., Karkera, J.D., dela Cruz, J., Roessler, E. and Muenke, M. (2002) 'CFC1 mutations in patients with transposition of the great arteries and double-outlet right ventricle', *Am J Hum Genet*, 70(3), pp. 776-80.

Gonzaga-Jauregui, C., Lupski, J.R. and Gibbs, R.A. (2012) 'Human genome sequencing in health and disease', *Annu Rev Med*, 63, pp. 35-61.

Goodship, J.A., Hall, D., Topf, A., Mamasoula, C., Griffin, H., Rahman, T.J., Glen, E., Tan, H., Palomino Doza, J., Relton, C.L., Bentham, J., Bhattacharya, S., Cosgrove, C., Brook, D., Granados-Riveron, J., Bu'Lock, F.A., O'Sullivan, J., Stuart, A.G., Parsons, J., Cordell, H.J. and Keavney, B. (2012) 'A Common Variant in the PTPN11 Gene Contributes to the Risk of Tetralogy of Fallot', *Circulation: Cardiovascular Genetics*, 5(3), pp. 287-292.

Goor, D.A., Dische, R. and Lillehei, C.W. (1972) 'The Conotruncus: I. Its Normal Inversion and Conus Absorption', *Circulation*, 46(2), pp. 375-384.

Goor, D.A. and Edwards, J.E. (1973) 'The Spectrum of Transposition of the Great Arteries: With Specific Reference to Developmental Anatomy of the Conus', *Circulation*, 48(2), pp. 406-415.

Gorini, F., Chiappa, E., Gargani, L. and Picano, E. (2014) 'Potential effects of environmental chemical contamination in congenital heart disease', *Pediatr Cardiol*, 35(4), pp. 559-68.

Greenway, S.C., Pereira, A.C., Lin, J.C., DePalma, S.R., Israel, S.J., Mesquita, S.M., Ergul, E., Conta, J.H., Korn, J.M., McCarroll, S.A., Gorham, J.M., Gabriel, S., Altshuler, D.M., Quintanilla-Dieck, M.D., Artunduaga, M.A., Eavey, R.D., Plenge, R.M., Shadick, N.A., Weinblatt, M.E., De Jager, P.L., Hafler, D.A., Breitbart, R.E., Seidman, J.G. and Seidman, C.E. (2009) 'De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot', *Nature Genetics*, 41(8), pp. 931-U98.

Griffin, H.R., Töpf, A., Glen, E., Zweier, C., Stuart, A.G., Parsons, J., Peart, I., Deanfield, J., O'Sullivan, J., Rauch, A., Scambler, P., Burn, J., Cordell, H.J., Keavney, B. and Goodship, J.A. (2010) '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*, 96(20), pp. 1651-1655.

Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., Duncan, L.E. and Borgwardt, K.M. (2015) 'The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity', *Human Mutation*, 36(5), pp. 513-523.

Grunert, M., Dorn, C., Schueler, M., Dunkel, I., Schlesinger, J., Mebus, S., Alexi-Meskishvili, V., Perrot, A., Wassilew, K., Timmermann, B., Hetzer, R., Berger, F. and Sperling, S.R. (2014) 'Rare and private variations in neural crest, apoptosis and sarcomere genes define the polygenic background of isolated Tetralogy of Fallot', *Human Molecular Genetics*, 23(12), pp. 3115-3128.

Guida, V., Ferese, R., Rocchetti, M., Bonetti, M., Sarkozy, A., Cecchetti, S., Gelmetti, V., Lepri, F., Copetti, M., Lamorte, G., Cristina Digilio, M., Marino, B., Zaza, A., den Hertog, J., Dallapiccola, B. and De Luca, A. (2013) 'A variant in the carboxyl-terminus of connexin 40 alters GAP junctions and increases risk for tetralogy of Fallot', *Eur J Hum Genet*, 21(1), pp. 69-75.

Guleserian, K.J. (2011) 'Adult congenital heart disease: surgical advances and options', *Prog Cardiovasc Dis*, 53(4), pp. 254-64.

Guo, Y., Ye, F., Sheng, Q., Clark, T. and Samuels, D.C. (2013) 'Three-stage quality control strategies for DNA re-sequencing data', *Briefings in Bioinformatics*.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonilla, E. and Martin, J.B. (1983) 'A polymorphic DNA marker genetically linked to Huntington's disease', *Nature*, 306(5940), pp. 234-238.

Haiko, P., Makinen, T., Keskitalo, S., Taipale, J., Karkkainen, M.J., Baldwin, M.E., Stacker, S.A., Achen, M.G. and Alitalo, K. (2008) 'Deletion of Vascular endothelial growth factor C (VEGF-C) and VEGF-D is not equivalent to VEGF receptor 3 deletion in mouse embryos', *Molecular and Cellular Biology*, 28(15), pp. 4843-4850.

Han, F. and Pan, W. (2010) 'A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants', *Human Heredity*, 70(1), pp. 42-54.

Hochstrasser, L., Ruchat, P., Sekarski, N., Hurni, M. and von Segesser, L.K. (2014) 'Long-term outcome of congenital aortic valve stenosis: predictors of reintervention', *Cardiology in the Young*, pp. 1-10.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. and McCombie, W.R. (2007) 'Genome-wide in situ exon capture for selective resequencing', *Nat Genet*, 39(12), pp. 1522-1527.

Hoffman, J.I. and Kaplan, S. (2002) 'The incidence of congenital heart disease', *J Am Coll Cardiol*, 39(12), pp. 1890-900.

Hoffman, J.I., Kaplan, S. and Liberthson, R.R. (2004) 'Prevalence of congenital heart disease', *Am Heart J*, 147(3), pp. 425-39.

Hoischen, A., Krumm, N. and Eichler, E.E. (2014) 'Prioritization of neurodevelopmental disease genes by discovery of new mutations', *Nat Neurosci*, 17(6), pp. 764-72.

Hoischen, A., van Bon, B.W.M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G., Devriendt, K., Amorim, M.Z., Revencu, N., Kidd, A., Barbosa, M., Turner, A., Smith, J., Oley, C., Henderson, A., Hayes, I.M., Thompson, E.M., Brunner, H.G., de Vries, B.B.A. and Veltman, J.A. (2010) 'De novo mutations of SETBP1 cause Schinzel-Giedion syndrome', *Nat Genet*, 42(6), pp. 483-485.

Houniet, D.T., Rahman, T.J., Al Turki, S., Hurles, M.E., Xu, Y., Goodship, J., Keavney, B. and Santibanez Koref, M. (2015) 'Using population data for assessing next-generation sequencing performance', *Bioinformatics*, 31(1), pp. 56-61.

Houslay, M.D., Baillie, G.S. and Maurice, D.H. (2007) 'cAMP-Specific phosphodiesterase-4 enzymes in the cardiovascular system: a molecular toolbox for generating compartmentalized cAMP signaling', *Circulation Research*, 100(7), pp. 950-66.

Hu, D., Barajas-Martinez, H., Pfeiffer, R., Dezi, F., Pfeiffer, J., Buch, T., Betzenhauser, M.J., Belardinelli, L., Kahlig, K.M., Rajamani, S., DeAntonio, H.J., Myerburg, R.J., Ito, H., Deshmukh, P., Marieb, M., Nam, G.B., Bhatia, A., Hasdemir, C., Haissaguerre, M., Veltmann, C., Schimpf, R., Borggrefe, M., Viskin, S. and Antzelevitch, C. (2014) 'Mutations in SCN10A Are Responsible for a Large Fraction of Cases of Brugada Syndrome', *Journal of the American College of Cardiology*, 64(1), pp. 66-79.

Huang, R.T., Xue, S., Xu, Y.J. and Yang, Y.Q. (2013) 'Somatic mutations in the GATA6 gene underlie sporadic tetralogy of Fallot', *Int J Mol Med*, 31(1), pp. 51-8.

Illumina (2012a) *CASAVA v.1.8.2 User Guide*. Available at: https://support.illumina.com/content/dam/illumina-support/documents/myillumina/a557afc4-bf0e-4dad-9e59-9c740dd1e751/casava_userguide_15011196d.pdf (Accessed: 8th of August).

Illumina (2012b) *Genome Analyzer IIx User Guide.* Available at: http://support.illumina.com/content/dam/illumina-support/documents/myillumina/d2aa31fa-51a0-48c9-8747-edfb748701ff/gaiix_userguide_scs2-10_15030966_c.pdf (Accessed: 8th of August).

Illumina (2014a) *HiSeq® 2000 System User Guide.* Available at: http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2000/hiseq-2000-user-guide-15011190-v.pdf (Accessed: 8th of August).

Illumina (2014b) *HiSeq® 2500 System User Guide*. Available at: http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2500/hiseq-2500-user-guide-15035786-d.pdf (Accessed: 12th of August).

Illumina (2015a) *Human660W-Quad BeadChip Support*. Available at: https://support.illumina.com/array/array_kits/human660w-quad_dna_analysis_kit.html (Accessed: 11th of August).

Illumina (2015b) *Paired-End Sequencing.* Available at: http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html (Accessed: 8th of August).

Illumina (2015c) *Sequencing and array-based solutions for genetic research.* Available at: http://www.illumina.com (Accessed: 8th of August).

Illumina (2015d) *Sequencing Software Support*. Available at: http://support.illumina.com/sequencing/sequencing_software.html (Accessed: 8th of August).

Iyengar, S.K. and Elston, R.C. (2007) 'The genetic basis of complex traits: rare variants or "common gene, common disease"?', *Methods Mol Biol*, 376, pp. 71-84.

Jenkins, K.J., Correa, A., Feinstein, J.A., Botto, L., Britt, A.E., Daniels, S.R., Elixson, M., Warnes, C.A. and Webb, C.L. (2007) '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*, 115(23), pp. 2995-3014.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. and von Mering, C. (2009) 'STRING 8—a global view on proteins and their functional interactions in 630 organisms', *Nucleic Acids Research*, 37(suppl 1), pp. D412-D416.

Jiang, J.Q., Li, R.G., Wang, J., Liu, X.Y., Xu, Y.J., Fang, W.Y., Chen, X.Z., Zhang, W., Wang, X.Z. and Yang, Y.Q. (2013) 'Prevalence and spectrum of GATA5 mutations associated with congenital heart disease', *Int J Cardiol*, 165(3), pp. 570-3.

Kaemmerer, H., Meisner, H., Hess, J. and Perloff, J.K. (2004) 'Surgical treatment of patent ductus arteriosus: a new historical perspective', *The American journal of cardiology*, 94(9), pp. 1153-1154.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*, 28(1), pp. 27-30.

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M. and Stenberg, P. (2015) 'Scaffolding of a bacterial genome using MinION nanopore sequencing', *Sci. Rep.*, 5.

Kaza, A., Minich, L.L. and Tani, L. (2013) 'Atrioventricular Septal Defects', in Da Cruz, E.M., Ivy, D. and Jaggers, J. (eds.) *Pediatric and Congenital Cardiology, Cardiac Surgery and Intensive Care*. Springer London, pp. 1479-1491.

Keinan, A. and Clark, A.G. (2012) 'Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants', *Science*, 336(6082), pp. 740-743.

Kendall, S., Karamichalis, J., Karamlou, T., Teitel, D. and Cohen, G. (2014) 'Atrial Septal Defect', *Pediatric and Congenital Cardiology, Cardiac Surgery and Intensive Care*, pp. 1439-1454.

Khumalo, N.P., Pillay, K., Beighton, P., Wainwright, H., Walker, B., Saxe, N., Mayosi, B.M. and Bateman, E.D. (2006) 'Poikiloderma, tendon contracture and pulmonary fibrosis: a new autosomal dominant syndrome?', *British Journal of Dermatology*, 155(5), pp. 1057-1061.

Kim, I., Moon, S.O., Koh, K.N., Kim, H., Uhm, C.S., Kwak, H.J., Kim, N.G. and Koh, G.Y. (1999) 'Molecular cloning, expression, and characterization of angiopoietin-related protein - Angiopoietin-related protein induces endothelial cell sprouting', *Journal of Biological Chemistry*, 274(37), pp. 26523-26528.

Kircher, M., Heyn, P. and Kelso, J. (2011) 'Addressing challenges in the production and analysis of illumina sequencing data', *BMC Genomics*, 12, p. 382.

Koboldt, D.C., Zhang, Q.Y., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K. (2012) 'VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing', *Genome Research*, 22(3), pp. 568-576.

Kobrynski, L.J. and Sullivan, K.E. (2007) 'Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes', *Lancet*, 370(9596), pp. 1443-1452.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U. and Stefansson, K. (2012) 'Rate of de novo mutations, father's age, and disease risk', *Nature*, 488(7412), pp. 471-475.

Koren, S. and Phillippy, A.M. (2015) 'One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly', *Curr Opin Microbiol*, 23, pp. 110-20.

Korpelainen, E.I., Karkkainen, M., Gunji, Y., Vikkula, M. and Alitalo, K. (1999) 'Endothelial receptor tyrosine kinases activate the STAT signaling pathway: mutant Tie-2 causing venous malformations signals a distinct STAT activation response', *Oncogene*, 18(1), pp. 1-8.

Ku, C.S., Vasiliou, V. and Cooper, D.N. (2012) 'A new era in the discovery of de novo mutations underlying human genetic disease', *Human Genomics*, 6.

Kukk, E., Lymboussaki, A., Taira, S., Kaipainen, A., Jeltsch, M., Joukov, V. and Alitalo, K. (1996) 'VEGF-C receptor binding and pattern of expression with VEGFR-3 suggests a role in lymphatic vascular development', *Development*, 122(12), pp. 3829-3837.

Kung, G. and Wong, P. (2014) 'Ventricular Septal Defects', in Wong, P.C. and Miller-Hance, W.C. (eds.) *Transesophageal Echocardiography for Congenital Heart Disease*. Springer London, pp. 241-252.

Lahdenranta, J., Hagendoorn, J., Padera, T.P., Hoshida, T., Nelson, G., Kashiwagi, S., Jain, R.K. and Fukumura, D. (2009) 'Endothelial Nitric Oxide Synthase Mediates Lymphangiogenesis and Lymphatic Metastasis', *Cancer Research*, 69(7), pp. 2801-2808.

Lange, K. (1997) *Mathematical and statistical methods for genetic analysis*. New York: Springer.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biol*, 10(3), p. R25.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K. and Ding, L. (2012) 'SomaticSniper: identification of somatic point mutations in whole genome sequencing data', *Bioinformatics*, 28(3), pp. 311-317.

LeCouter, J., Lin, R., Tejada, M., Frantz, G., Peale, F., Hillan, K.J. and Ferrara, N. (2003) 'The endocrine-gland-derived VEGF homologue Bv8 promotes angiogenesis in the testis: Localization of Bv8 receptors to endothelial cells', *Proceedings of the National Academy of Sciences of the United States of America*, 100(5), pp. 2685-2690.

Ledergerber, C. and Dessimoz, C. (2011) 'Base-calling for next-generation sequencing platforms', *Briefings in Bioinformatics*.

Lee, M., d'Udekem, Y. and Brizard, C. (2014a) 'Coarctation of the Aorta', *Pediatric and Congenital Cardiology, Cardiac Surgery and Intensive Care*, pp. 1631-1646.

Lee, S., Abecasis, G.R., Boehnke, M. and Lin, X. (2014b) 'Rare-variant association analysis: study designs and statistical tests', *American Journal of Human Genetics*, 95(1), pp. 5-23.

Leeds, J.S., Hopper, A.D. and Sanders, D.S. (2008) 'Coeliac disease', *British Medical Bulletin*, 88(1), pp. 157-170.

Lehtokari, V.-L., Pelin, K., Sandbacka, M., Ranta, S., Donner, K., Muntoni, F., Sewry, C., Angelini, C., Bushby, K., Van den Bergh, P., Iannaccone, S., Laing, N.G. and Wallgren-Pettersson, C. (2006) 'Identification of 45 novel mutations in the nebulin gene associated with autosomal recessive nemaline myopathy', *Human Mutation*, 27(9), pp. 946-956.

Lev, M. and Eckner, F.A.O. (1964) 'THe pathologic anatomy of tetralogy of fallot and its variations', *Chest*, 45(3), pp. 251-261.

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M. and Abecasis, G.R. (2012) 'A likelihood-based framework for variant calling and de novo mutation detection in families', *PLoS Genet*, 8(10), p. e1002944.

Li, B. and Leal, S.M. (2008) 'Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data', *The American Journal of Human Genetics*, 83(3), pp. 311-321.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27(21), pp. 2987-2993.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009a) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078-2079.

Li, H. and Homer, N. (2010) 'A survey of sequence alignment algorithms for next-generation sequencing', *Brief Bioinform*, 11(5), pp. 473-83.

Li, H., Ruan, J. and Durbin, R. (2008) 'Mapping short DNA sequencing reads and calling variants using mapping quality scores', *Genome Res*, 18(11), pp. 1851-8.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009b) 'SOAP2: an improved ultrafast tool for short read alignment', *Bioinformatics*, 25(15), pp. 1966-7.

Li, Y., Klena, N.T., Gabriel, G.C., Liu, X.Q., Kim, A.J., Lemke, K., Chen, Y., Chatterjee, B., Devine, W., Damerla, R.R., Chang, C.F., Yagi, H., San Agustin, J.T., Thahir, M., Anderton, S., Lawhead, C., Vescovi, A., Pratt, H., Morgan, J., Haynes, L., Smith, C.L., Eppig, J.T., Reinholdt, L., Francis, R., Leatherbury, L., Ganapathiraju, M.K., Tobita, K., Pazour, G.J. and Lo, C.W. (2015) 'Global genetic analysis in mice unveils central role for cilia in congenital heart disease', *Nature*, 521(7553), pp. 520-U224.

Liberfarb, R.M., Goldblatt, A., Opitz, J.M. and Reynolds, J.F. (1986) 'Prevalence of mitral-valve prolapse in the Stickler syndrome', *American Journal of Medical Genetics*, 24(3), pp. 387-392.

Liu, S., Joseph, K.S., Lisonkova, S., Rouleau, J., Van den Hof, M., Sauve, R. and Kramer, M.S. (2013) 'Association between maternal chronic conditions and congenital heart defects: a population-based cohort study', *Circulation*, 128(6), pp. 583-9.

Loffredo, C.A., Silbergeld, E.K., Ferencz, C. and Zhang, J. (2001) 'Association of transposition of the great arteries in infants with maternal exposures to herbicides and rodenticides', *Am J Epidemiol*, 153(6), pp. 529-36.

London Medical Databases (2014) *London Medical Databases: About the LM Database Series.* Available at: http://www.lmdatabases.com/about_lmd.html#lddb (Accessed: 14th of September).

Longo, N., Amat di San Filippo, C. and Pasquali, M. (2006) 'Disorders of carnitine transport and the carnitine cycle', *American journal of medical genetics. Part C, Seminars in medical genetics*, 142C(2), pp. 77-85.

Lui, V.C.H., Ng, L.J., Sat, E.W.Y. and Cheah, K.S.E. (1996) 'The human alpha 2(XI) collagen gene (COL11A2): Completion of coding information, identification of the promoter sequence, and precise localization within the major histocompatibility complex reveal overlap with the KE5 gene', *Genomics*, 32(3), pp. 401-412.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., Barrett, J.C., Biesecker, L.G., Conrad, D.F., Cooper, G.M., Cox, N.J., Daly, M.J., Gerstein, M.B., Goldstein, D.B., Hirschhorn, J.N., Leal, S.M., Pennacchio, L.A., Stamatoyannopoulos, J.A., Sunyaev, S.R., Valle, D., Voight, B.F., Winckler, W.

and Gunter, C. (2014) 'Guidelines for investigating causality of sequence variants in human disease', *Nature*, 508(7497), pp. 469-476.
Madsen, B.E. and Browning, S.R. (2009) 'A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic', *PLoS Genetics*, 5(2), p. e1000384.
Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. and Jabado, N. (2011) 'What can exome sequencing do for you?', *Journal of Medical Genetics*, 48(9), pp. 580-589.
Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) 'Target-enrichment strategies for next-generation sequencing', *Nat Meth*, 7(2), pp. 111-118.
Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A. and Visscher, P.M. (2009) 'Finding the missing heritability of complex diseases', *Nature*, 461(7265), pp. 747-53.
Mardis, E.R. (2013) 'Next-generation sequencing platforms', *Annu Rev Anal Chem (Palo Alto Calif)*, 6, pp. 287-303.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437(7057), pp. 376-80.
Marth, G.T., Yu, F.L., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., Chen, Y., Challis, D., Clarke, L., Ball, E.V., Cibulskis, K., Cooper, D.N., Fulton, B., Hartl, C., Koboldt, D., Muzny, D., Smith, R., Sougnez, C., Stewart, C., Ward, A., Yu, J., Xue, Y.L., Altshuler, D., Bustamante, C.D., Clark, A.G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A., Gibbs, R. and Project, G. (2011) 'The functional spectrum of low-frequency coding variation', *Genome Biology*, 12(9).
Martins, P. and Castela, E. (2008) 'Transposition of the great arteries', *Orphanet Journal of Rare Diseases*, 3.
Massingham, T. and Goldman, N. (2012) 'All Your Base: a fast and accurate probabilistic approach to base calling', *Genome Biology*, 13(2), p. R13.
Mayr, J.A., Merkel, O., Kohlwein, S.D., Gebhardt, B.R., Bohles, H., Fotschl, U., Koch, J., Jaksch, M., Lochmuller, H., Horvath, R., Freisinger, P. and Sperl, W. (2007) 'Mitochondrial phosphate-carrier deficiency: A novel disorder of oxidative phosphorylation', *American Journal of Human Genetics*, 80(3), pp. 478-484.
Mazur, W., Siegel, M.J., Miszalski-Jamka, T. and Pelberg, R. (2013) 'Tetralogy of Fallot Repair', in *CT Atlas of Adult Congenital Heart Disease*. Springer, pp. 311-318.
McElhinney, D.B., Geiger, E., Blinder, J., Benson, D.W. and Goldmuntz, E. (2003) 'NKX2.5 mutations in patients with congenital heart disease', *J Am Coll Cardiol*, 42(9), pp. 1650-5.
McElhinney, D.B., Krantz, I.D., Bason, L., Piccoli, D.A., Emerick, K.M., Spinner, N.B. and Goldmuntz, E. (2002) 'Analysis of cardiovascular phenotype and genotype-phenotype correlation in individuals with a JAG1 mutation and/or Alagille syndrome', *Circulation*, 106(20), pp. 2567-2574.
Mercier, S., Kury, S., Shaboodien, G., Houniet, D.T., Khumalo, N.P., Bou-Hanna, C., Bodak, N., Cormier-Daire, V., David, A., Faivre, L., Figarella-Branger, D., Gherardi, R.K., Glen, E., Hamel, A., Laboisse, C., Le Caignec, C., Lindenbaum, P., Magot, A., Munnich, A., Mussini, J.M., Pillay, K., Rahman, T., Redon, R., Salort-Campana, E., Santibanez-Koref, M., Thauvin, C., Barbarot, S.,

Keavney, B., Bezieau, S. and Mayosi, B.M. (2013) 'Mutations in FAM111B Cause Hereditary Fibrosing Poikiloderma with Tendon Contracture, Myopathy, and Pulmonary Fibrosis', *American Journal of Human Genetics*, 93(6), pp. 1100-1107.

Metzker, M.L. (2010) 'Sequencing technologies - the next generation', *Nat Rev Genet*, 11(1), pp. 31-46.

Meyerson, M., Gabriel, S. and Getz, G. (2010) 'Advances in understanding cancer genomes through second-generation sequencing', *Nature Reviews Genetics*, 11(10), pp. 685-696.

Michaelovsky, E., Frisch, A., Carmel, M., Patya, M., Zarchi, O., Green, T., Basel-Vanagaite, L., Weizman, A. and Gothelf, D. (2012) 'Genotype-phenotype correlation in 22q11.2 deletion syndrome', *BMC Medical Genetics*, 13, pp. 122-122.

Michelucci, A., Ghirri, P., Iacopetti, P., Conidi, M.E., Fogli, A., Baldinotti, F., Lunardi, S., Forli, F., Moscuzza, F., Berrettini, S., Boldrini, A., Simi, P. and Pellegrini, S. (2010) 'Identification of three novel mutations in the CHD7 gene in patients with clinical signs of typical or atypical CHARGE syndrome', *Int J Pediatr Otorhinolaryngol*, 74(12), pp. 1441-4.

Mitchell, S.C., Korones, S.B. and Berendes, H.W. (1971) 'Congenital heart disease in 56,109 births. Incidence and natural history', *Circulation*, 43(3), pp. 323-32.

Morgenthaler, S. and Thilly, W.G. (2007) 'A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1–2), pp. 28-56.

Morris, A.P. and Zeggini, E. (2010) 'An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies', *Genetic Epidemiology*, 34(2), pp. 188-193.

Moutsianas, L. and Morris, A.P. (2014) 'Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits', *Briefings in Functional Genomics*, 13(5), pp. 362-370.

Nash, D., Arrington, C.B., Kennedy, B.J., Yandell, M., Wu, W., Zhang, W., Ware, S., Jorde, L.B., Gruber, P.J., Yost, H.J., Bowles, N.E. and Bleyl, S.B. (2015) 'Shared Segment Analysis and Next-Generation Sequencing Implicates the Retinoic Acid Signaling Pathway in Total Anomalous Pulmonary Venous Return (TAPVR)', *PLoS One*, 10(6), p. e0131514.

NCBI (2015) *Online Mendelian Inheritance in Man, OMIM®*. Available at: http://www.omim.org/.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K. and Daly, M.J. (2011) 'Testing for an Unusual Distribution of Rare Variants', *PLoS Genetics*, 7(3).

Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) 'Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes', *Science*, 324(5925), pp. 387-389.

Nelson, J.S., Bove, E.L. and Hirsch-Romano, J.C. (2014) 'Tetralogy of Fallot', *Pediatric and Congenital Cardiology, Cardiac Surgery and Intensive Care*, pp. 1505-1526.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., Jean, P.S., Verzilli, C., Shen, J.D., Tang, Z.Z., Bacanu, S.A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M.D., Nangle, K., Wang, J., Abecasis, G., Cardon, L.R., Zollner, S., Whittaker, J.C., Chissoe, S.L., Novembre, J. and Mooser, V. (2012) 'An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People', *Science*, 337(6090), pp. 100-104.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A. and Shendure, J. (2009) 'Targeted capture and massively parallel sequencing of 12 human exomes', *Nature*, 461(7261), pp. 272-6.

NHLBI (2015) *Exome Variant Server*. Available at: http://evs.gs.washington.edu/EVS/ (Accessed: 8th of August).

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) 'Genotype and SNP calling from next-generation sequencing data', *Nat Rev Genet*, 12(6), pp. 443-51.

Niessen, K. and Karsan, A. (2008) 'Notch signaling in cardiac development', *Circulation Research*, 102(10), pp. 1169-1181.

Nishiguchi, K.M., Tearle, R.G., Liu, Y.P., Oh, E.C., Miyake, N., Benaglio, P., Harper, S., Koskiniemi-Kuendig, H., Venturini, G., Sharon, D., Koenekoop, R.K., Nakamura, M., Kondo, M., Ueno, S., Yasuma, T.R., Beckmann, J.S., Ikegawa, S., Matsumoto, N., Terasaki, H., Berson, E.L., Katsanis, N. and Rivolta, C. (2013) 'Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene', *Proceedings of the National Academy of Sciences*, 110(40), pp. 16139-16144.

Nora, J.J. (1968) 'Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction', *Circulation*, 38(3), pp. 604-17.

Novocraft (2014a) *FAQ | Novocraft*. Available at: http://www.novocraft.com/support/faq/.

Novocraft (2014b) *Novoalign.* Available at: http://www.novocraft.com/products/novoalign/ (Accessed: 8th of August).

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K. and Lyon, G. (2013) 'Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing', *Genome Medicine*, 5(3), p. 28.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J.B., Turner, E.H., Levy, R., O'Day, D.R., Krumm, N., Coe, B.P., Martin, B.K., Borenstein, E., Nickerson, D.A., Mefford, H.C., Doherty, D., Akey, J.M., Bernier, R., Eichler, E.E. and Shendure, J. (2012a) 'Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders', *Science*, 338(6114), pp. 1619-22.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., Turner, E.H., Stanaway, I.B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J.M., Borenstein, E., Rieder, M.J., Nickerson, D.A., Bernier, R., Shendure, J. and Eichler, E.E. (2012b) 'Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations', *Nature*, 485(7397), pp. 246-50.

Obler, D., Juraszek, A.L., Smoot, L.B. and Natowicz, M.R. (2008) 'Double outlet right ventricle: aetiologies and associations', *Journal of Medical Genetics*, 45(8), pp. 481-497.

Open Grid Scheduler project (2013) *Open Grid Scheduler: The official Open Source Grid Engine*. Available at: http://gridscheduler.sourceforge.net/index.html.

Oyen, N., Poulsen, G., Boyd, H.A., Wohlfahrt, J., Jensen, P.K. and Melbye, M. (2009) 'Recurrence of congenital heart defects in families', *Circulation*, 120(4), pp. 295-301.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z. (2014) 'A survey of tools for variant analysis of next-generation genome sequencing data', *Briefings in Bioinformatics*, 15(2), pp. 256-278.

Park, M.S., Ludwig, D.L., Stigger, E. and Lee, S.H. (1996) 'Physical interaction between human RAD52 and RPA is required for homologous recombination in mammalian cells', *Journal of Biological Chemistry*, 271(31), pp. 18996-9000.

Patel, R.K. and Jain, M. (2012) 'NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data', *PLoS One*, 7(2).

Peng, G., Fan, Y., Palculict, T.B., Shen, P., Ruteshouser, E.C., Chi, A.K., Davis, R.W., Huff, V., Scharfe, C. and Wang, W. (2013) 'Rare variant detection using family-based sequencing analysis', *Proc Natl Acad Sci U S A*, 110(10), pp. 3985-90.

Penny, D.J. and Vick, G.W., 3rd (2011) 'Ventricular septal defect', *Lancet*, 377(9771), pp. 1103-12.

Perl.org (2002) *The Perl Programming Language.* (Accessed: 8th of August).

Perloff, J.K. and Warnes, C.A. (2001) 'Challenges posed by adults with repaired congenital heart disease', *Circulation*, 103(21), pp. 2637-43.

Pierpont, M.E., Basson, C.T., Benson, D.W., Jr., Gelb, B.D., Giglia, T.M., Goldmuntz, E., McGee, G., Sable, C.A., Srivastava, D. and Webb, C.L. (2007) '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*, 115(23), pp. 3015-38.

Pizzuti, A., Sarkozy, A., Newton, A.L., Conti, E., Flex, E., Digilio, M.C., Amati, F., Gianni, D., Tandoi, C., Marino, B., Crossley, M. and Dallapiccola, B. (2003) 'Mutations of ZFPM2/FOG2 gene in sporadic cases of tetralogy of Fallot', *Hum Mutat*, 22(5), pp. 372-7.

Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W., Hambleton, S., Burns, S.O., Thrasher, A.J., Kumararatne, D., Doffinger, R. and Nejentsev, S. (2012) 'A robust model for read count data in exome sequencing experiments and implications for copy number variant calling', *Bioinformatics*, 28(21), pp. 2747-2754.

Prieto, L.R. (2005) 'Management of Tetralogy of Fallot with Pulmonary Atresia', *Images in Paediatric Cardiology*, 7(3), pp. 24-42.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D. and Ostell, J.M. (2014) 'RefSeq: an update on mammalian reference sequences', *Nucleic Acids Research*, 42(D1), pp. D756-D763.

QIAGEN (2015) *Ingenuity Variant Analysis™*. Available at: www.quiagen.com/ingenuity (Accessed: 2nd of September).

Rabbani, B., Tekin, M. and Mahdieh, N. (2014) 'The promise of whole-exome sequencing in medical genetics', *J Hum Genet*, 59(1), pp. 5-15.

Rae, J.M., Johnson, M.D., Cordero, K.E., Scheys, J.O., Larios, J.M., Gottardis, M.M., Pienta, K.J. and Lippman, M.E. (2006) 'GREB1 is a novel androgen-regulated gene required for prostate cancer growth', *Prostate*, 66(8), pp. 886-94.

Raff, G.W., Geiss, D.M., Shah, J.J., Bond, L.M. and Carroll, J.A. (2002) 'Repair of transposition of the great arteries with total anomalous pulmonary venous return', *Ann Thorac Surg*, 73(2), pp. 655-7.

Ramu, A., Noordam, M.J., Schwartz, R.S., Wuster, A., Hurles, M.E., Cartwright, R.A. and Conrad, D.F. (2013) 'DeNovoGear: de novo indel and point mutation discovery and phasing', *Nat Methods*, 10(10), pp. 985-7.

Ranade, S.S., Qiu, Z.Z., Woo, S.H., Hur, S.S., Murthy, S.E., Cahalan, S.M., Xu, J., Mathur, J., Bandell, M., Coste, B., Li, Y.S.J., Chien, S. and Patapoutian, A. (2014) 'Piezo1, a mechanically activated ion channel, is required for vascular development in mice', *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), pp. 10347-10352.

Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Hüffmeier, U., Thiel, C., Rüschendorf, F., Nürnberg, P., Reis, A. and Trautmann, U. (2006) 'Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation', *American Journal of Medical Genetics Part A*, 140A(19), pp. 2063-2074.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998) 'GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support', *Bioinformatics*, 14(8), pp. 656-664.

Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein mutations: application to cancer genomics', *Nucleic Acids Research*.

Richards, A.A. and Garg, V. (2010) 'Genetics of congenital heart disease', *Current cardiology reviews*, 6(2), pp. 91-97.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L. and et al. (1989) 'Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA', *Science*, 245(4922), pp. 1066-73.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., Ferreira, P.G., Smith, K.S., Zhang, R., Zhao, F.M., Banks, E., Poplin, R., Ruderfer, D.M., Purcell, S.M., Tukiainen, T., Minikel, E.V., Stenson, P.D., Cooper, D.N., Huang,

K.H., Sullivan, T.J., Nedzel, J., Bustamante, C.D., Li, J.B., Daly, M.J., Guigo, R., Donnelly, P., Ardlie, K., Sammeth, M., Dermitzakis, E.T., McCarthy, M.I., Montgomery, S.B., Lappalainen, T., MacArthur, D.G., Consortium, G. and Consortium, G. (2015) 'Effect of predicted protein-truncating genetic variants on the human transcriptome', *Science*, 348(6235), pp. 666-669.
Robasky, K., Lewis, N.E. and Church, G.M. (2014) 'The role of replicates for error mitigation in next-generation sequencing', *Nat Rev Genet*, 15(1), pp. 56-62.
Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) 'Integrative genomics viewer', *Nat Biotech*, 29(1), pp. 24-26.
Robinson, P.N. (2010) 'Whole-exome sequencing for finding de novo mutations in sporadic mental retardation', *Genome Biology*, 11(12).
Rochais, F., Dandonneau, M., Mesbah, K., Jarry, T., Mattei, M.G. and Kelly, R.G. (2009) 'Hes1 Is Expressed in the Second Heart Field and Is Required for Outflow Tract Development', *PLoS One*, 4(7).
Roessler, E., Ouspenskaia, M.V., Karkera, J.D., Velez, J.I., Kantipong, A., Lacbawan, F., Bowers, P., Belmont, J.W., Towbin, J.A., Goldmuntz, E., Feldman, B. and Muenke, M. (2008) 'Reduced NODAL signaling strength via mutation of several pathway members including FOXH1 is linked to human heart defects and holoprosencephaly', *Am J Hum Genet*, 83(1), pp. 18-29.
Roessler, E., Pei, W.H., Ouspenskaia, M.V., Karkera, J.D., Velez, J.I., Banerjee-Basu, S., Gibney, G., Lupo, P.J., Mitchell, L.E., Towbin, J.A., Bowers, P., Belmont, J.W., Goldmuntz, E., Baxevanis, A.D., Feldman, B. and Muenke, M. (2009) 'Cumulative ligand activity of NODAL mutations and modifiers are linked to human heart defects and holoprosencephaly', *Molecular Genetics and Metabolism*, 98(1-2), pp. 225-234.
Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C. and Jaffe, D.B. (2013) 'Characterizing and measuring bias in sequence data', *Genome Biology*, 14(5), p. R51 [Online]. Available at: http://europepmc.org/abstract/MED/23718773

http://europepmc.org/articles/PMC4053816?pdf=render

http://europepmc.org/articles/PMC4053816

http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=EBI&pubmedid=23718773

http://www.pubmedcentral.nih.gov/picrender.fcgi?tool=EBI&pubmedid=23718773&action=stream&blobtype=pdf

http://dx.doi.org/10.1186/gb-2013-14-5-r51 DOI: 10.1186/gb-2013-14-5-r51 (Accessed: 2013).
Sadeghpour, A. and Alizadehasl, A. (2014) 'Pulmonary Valve Stenosis', in *Comprehensive Approach to Adult Congenital Heart Disease*. Springer, pp. 311-314.
Saint Pierre, A. and Genin, E. (2014) 'How important are rare variants in common disease?', *Brief Funct Genomics*, 13(5), pp. 353-361.
Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., Wall, D.P., MacArthur, D.G., Gabriel, S.B., DePristo, M., Purcell, S.M., Palotie, A., Boerwinkle, E., Buxbaum, J.D., Cook, E.H., Jr., Gibbs, R.A., Schellenberg, G.D., Sutcliffe, J.S., Devlin, B., Roeder, K., Neale, B.M. and Daly, M.J. (2014) 'A framework for the interpretation of de novo mutation in human disease', *Nat Genet*, 46(9), pp. 944-50.
Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., Walker, M.F., Ober, G.T., Teran, N.A., Song, Y., El-Fishawy, P., Murtha, R.C., Choi, M., Overton, J.D., Bjornson, R.D., Carriero, N.J., Meyer, K.A., Bilguvar, K., Mane, S.M., Sestan, N., Lifton, R.P., Gunel, M., Roeder, K., Geschwind, D.H., Devlin, B. and State, M.W. (2012) 'De novo mutations revealed by whole-exome sequencing are strongly associated with autism', *Nature*, 485(7397), pp. 237-U124.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA Sequencing with Chain-Terminating Inhibitors', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463-5467.

Saremi, F. (2014) 'Transposition of the Great Arteries', in Saremi, F. (ed.) *Cardiac CT and MR for Adult Congenital Heart Disease*. Springer New York, pp. 225-258.

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J. and Cheetham, R.K. (2012) 'Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs', *Bioinformatics*, 28(14), pp. 1811-1817.

Schadt, E.E., Turner, S. and Kasarskis, A. (2011) 'A window into third generation sequencing (vol 19, pg R227, 2010)', *Human Molecular Genetics*, 20(4), pp. 853-853.

Schneider, D.J. and Moore, J.W. (2006) 'Patent ductus arteriosus', *Circulation*, 114(17), pp. 1873-82.

Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. (2009) 'Common vs. rare allele hypotheses for complex diseases', *Curr Opin Genet Dev*, 19(3), pp. 212-9.

Schott, J.J., Benson, D.W., Basson, C.T., Pease, W., Silberbach, G.M., Moak, J.P., Maron, B.J., Seidman, C.E. and Seidman, J.G. (1998) 'Congenital heart disease caused by mutations in the transcription factor NKX2-5', *Science*, 281(5373), pp. 108-11.

Schultz, D.W., Klein, M.L., Humpert, A.J., Luzier, C.W., Persun, V., Schain, M., Mahan, A., Runckel, C., Cassera, M., Vittal, V., Doyle, T.M., Martin, T.M., Weleber, R.G., Francis, P.J. and Acott, T.S. (2003) 'Analysis of the ARMD1 locus: evidence that a mutation in HEMICENTIN-1 is associated with age-related macular degeneration in a large family', *Hum Mol Genet*, 12(24), pp. 3315-23.

Schwarz, J.M., Rodelsperger, C., Schuelke, M. and Seelow, D. (2010) 'MutationTaster evaluates disease-causing potential of sequence alterations', *Nat Meth*, 7(8), pp. 575-576.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) 'dbSNP: the NCBI database of genetic variation', *Nucleic Acids Res*, 29(1), pp. 308-11.

Shinebourne, E.A., Anderson, R.H. and Bowyer, J.J. (1975) 'Variations in clinical presentation of Fallot's tetralogy in infancy. Angiographic and pathogenetic implications', *British Heart Journal*, 37(9), pp. 946-955.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W. and Hajibabaei, M. (2014) 'Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens', *Molecular Ecology Resources*, 14(5), pp. 892-901.

Silversides, C.K., Lionel, A.C., Costain, G., Merico, D., Migita, O., Liu, B., Yuen, T., Rickaby, J., Thiruvahindrapuram, B., Marshall, C.R., Scherer, S.W. and Bassett, A.S. (2012) 'Rare Copy Number Variations in Adults with Tetralogy of Fallot Implicate Novel Risk Gene Pathways', *PLoS Genetics*, 8(8).

Sisakian, H. (2014) 'Cardiomyopathies: Evolution of pathogenesis concepts and potential for new therapies', *World Journal of Cardiology*, 6(6), pp. 478-494.

Siva, N. (2015) 'UK gears up to decode 100 000 genomes from NHS patients', *The Lancet*, 385(9963), pp. 103-104.

Sobreira, N.L.M., Cirulli, E.T., Avramopoulos, D., Wohler, E., Oswald, G.L., Stevens, E.L., Ge, D.L., Shianna, K.V., Smith, J.P., Maia, J.M., Gumbs, C.E., Pevsner, J., Thomas, G., Valle, D., Hoover-Fong, J.E. and Goldstein, D.B. (2010) 'Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene', *Plos Genetics*, 6(6).

Soemedi, R., Topf, A., Wilson, I.J., Darlay, R., Rahman, T., Glen, E., Hall, D., Huang, N., Bentham, J., Bhattacharya, S., Cosgrove, C., Brook, J.D., Granados-Riveron, J., Setchfield, K., Bu'Lock, F., Thornborough, C., Devriendt, K., Breckpot, J., Hofbeck, M., Lathrop, M., Rauch, A., Blue, G.M., Winlaw, D.S., Hurles, M., Santibanez-Koref, M., Cordell, H.J., Goodship, J.A. and Keavney, B.D. (2012a) 'Phenotype-specific effect of chromosome 1q21.1 rearrangements and GJA5 duplications in 2436 congenital heart disease patients and 6760 controls', *Human Molecular Genetics*, 21(7), pp. 1513-1520.

Soemedi, R., Wilson, I.J., Bentham, J., Darlay, R., Topf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., Blue, G.M., Breckpot, J., Hellens, S., Zwolinkski, S., Glen, E., Mamasoula, C., Rahman, T.J., Hall, D., Rauch, A., Devriendt, K., Gewillig, M., O' Sullivan, J., Winlaw, D.S., Bu'Lock, F., Brook, J.D., Bhattacharya, S., Lathrop, M., Santibanez-Koref, M., Cordell, H.J., Goodship, J.A. and Keavney, B.D. (2012b) 'Contribution of Global Rare Copy-Number Variants to the Risk of Sporadic Congenital Heart Disease', *American Journal of Human Genetics*, 91(3), pp. 489-501.

Stalmans, I., Lambrechts, D., De smet, F., Jansen, S., Wang, J., Maity, S., Kneer, P., von der Ohe, M., Swillen, A., Maes, C., Gewillig, M., Molin, D.G.M., Hellings, P., Boetel, T., Haardt, M., Compernolle, V., Dewerchin, M., Plaisance, S., Vlietinck, R., Emanuel, B., Gittenberger-de Groot, A.C., Scambler, P., Morrow, B., Driscol, D.A., Moons, L., Esguerra, C.V., Carmeliet, G., Behn-Krappa, A., Devriendt, K., Collen, D., Conway, S.J. and Carmeliet, P. (2003) 'VEGF: A modifier of the del22q11 (DiGeorge) syndrome?', *Nature Medicine*, 9(2), pp. 173-182.

Starr, J.P. (2010) 'Tetralogy of Fallot: Yesterday and Today', *World Journal of Surgery*, 34(4), pp. 658-668.

Stittrich, A.B., Lehman, A., Bodian, D.L., Ashworth, J., Zong, Z.Y., Li, H., Lam, P., Khromykh, A., Iyer, R.K., Vockley, J.G., Baveja, R., Silva, E.S., Dixon, J., Leon, E.L., Solomon, B.D., Glusman, G., Niederhuber, J.E., Roach, J.C. and Patel, M.S. (2014) 'Mutations in NOTCH1 Cause Adams-Oliver Syndrome', *American Journal of Human Genetics*, 95(3), pp. 275-284.

Tazume, H., Miyata, K., Tian, Z., Endo, M., Horiguchi, H., Takahashi, O., Horio, E., Tsukano, H., Kadomatsu, T., Nakashima, Y., Kunitomo, R., Kaneko, Y., Moriyama, S., Sakaguchi, H., Okamoto, K., Hara, M., Yoshinaga, T., Yoshimura, K., Aoki, H., Araki, K., Hao, H., Kawasuji, M. and Oike, Y. (2012) 'Macrophage-Derived Angiopoietin-Like Protein 2 Accelerates Development of Abdominal Aortic Aneurysm', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 32(6), pp. 1400-1409.

Teer, J.K. and Mullikin, J.C. (2010) 'Exome sequencing: the sweet spot before whole genomes', *Hum Mol Genet*, 19(R2), pp. R145-51.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D., Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J. and Akey, J.M. (2012) 'Evolution and functional impact of rare coding variation from deep sequencing of human exomes', *Science*, 337(6090), pp. 64-9.

The 1000 Genomes Project Consortium (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), pp. 56-65.

The Broad Institute (2015a) *Exome Aggregation Consortium (ExAC)*. Available at: http://exac.broadinstitute.org.

The Broad Institute (2015b) *GATK | Tool Documentation Index.* Available at: https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php#--standard_min_confidence_threshold_for_emitting (Accessed: 8th of August).

The Broad Institute (2015c) *Picard tools.* Available at: http://broadinstitute.github.io/picard/ (Accessed: 8th of August).

Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', *Brief Bioinform*, 14(2), pp. 178-92.

Tomita-Mitchell, A., Maslen, C.L., Morris, C.D., Garg, V. and Goldmuntz, E. (2007) 'GATA4 sequence variants in patients with congenital heart disease', *J Med Genet*, 44(12), pp. 779-83.

Topf, A., Griffin, H.R., Glen, E., Soemedi, R., Brown, D.L., Hall, D., Rahman, T.J., Eloranta, J.J., Jungst, C., Stuart, A.G., O'Sullivan, J., Keavney, B.D. and Goodship, J.A. (2014) 'Functionally significant, rare transcription factor variants in tetralogy of Fallot', *PLoS One*, 9(8), p. e95453.

UCSC (2015) *UCSC Genome Browser Home*. Available at: https://genome-euro.ucsc.edu/ (Accessed: 8th of August).

Unolt, M., Putotto, C., Silvestri, L.M., Marino, D., Scarabotti, A., Massaccesi, V., Caiaro, A., Versacci, P. and Marino, B. (2013) 'TRANSPOSITION OF GREAT ARTERIES: NEW INSIGHTS INTO THE PATHOGENESIS', *Frontiers in Pediatrics*, 1.

van der Linde, D., Konings, E.E., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J. and Roos-Hesselink, J.W. (2011) 'Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis', *J Am Coll Cardiol*, 58(21), pp. 2241-7.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014) 'Ten years of next-generation sequencing technology', *Trends in Genetics*, 30(9), pp. 418-426.

Veltman, J.A. and Brunner, H.G. (2012a) 'Applications of Next-Generation Sequencing De Novo Mutations in Human Genetic Disease', *Nature Reviews Genetics*, 13(8), pp. 565-575.

Veltman, J.A. and Brunner, H.G. (2012b) 'De novo mutations in human genetic disease', *Nat Rev Genet*, 13(8), pp. 565-75.

Vergales, J.E., Gangemi, J.J., Rhueban, K.S. and Lim, D.S. (2013) 'Coarctation of the aorta - the current state of surgical and transcatheter therapies', *Curr Cardiol Rev*, 9(3), pp. 211-9.

Verheugt, C.L., Uiterwaal, C.S., van der Velde, E.T., Meijboom, F.J., Pieper, P.G., van Dijk, A.P., Vliegen, H.W., Grobbee, D.E. and Mulder, B.J. (2010) 'Mortality in adult congenital heart disease', *Eur Heart J*, 31(10), pp. 1220-9.

Visscher, P.M., Hill, W.G. and Wray, N.R. (2008) 'Heritability in the genomics era [mdash] concepts and misconceptions', *Nat Rev Genet*, 9(4), pp. 255-266.

Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., van Bon, B.W.M., Hoischen, A., de Vries, B.B.A., Brunner, H.G. and Veltman, J.A. (2010) 'A de novo paradigm for mental retardation', *Nature Genetics*, 42(12), pp. 1109-+.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Res*, 38(16), p. e164.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K., Pao, W. and Zhao, Z. (2013) 'Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers', *Genome Medicine*, 5(10), p. 91.

Ware, S.M., Aygun, M.G. and Hildebrandt, F. (2011) 'Spectrum of clinical diseases caused by disorders of primary cilia', *Proc Am Thorac Soc*, 8(5), pp. 444-50.

Warnes, C.A. (2006) 'Transposition of the Great Arteries', *Circulation*, 114(24), pp. 2699-2709.

Warnes, C.A., Williams, R.G., Bashore, T.M., Child, J.S., Connolly, H.M., Dearani, J.A., del Nido, P., Fasules, J.W., Graham Jr, T.P., Hijazi, Z.M., Hunt, S.A., King, M.E., Landzberg, M.J., Miner, P.D., Radford, M.J., Walsh, E.P. and Webb, G.D. (2008) 'ACC/AHA 2008 Guidelines for the Management of Adults With Congenital Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Develop Guidelines on the Management of Adults With Congenital Heart Disease) Developed in Collaboration With the American Society of Echocardiography, Heart Rhythm Society, International Society for Adult Congenital Heart Disease, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons', *Journal of the American College of Cardiology*, 52(23), pp. e143-e263.

Wei, Z., Wang, W., Hu, P., Lyon, G.J. and Hakonarson, H. (2011) 'SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data', *Nucleic Acids Research*, 39, p. e132.

Wellcome Trust Sanger Institute (2010) *UK10K*. Available at: www.uk10k.org (Accessed: 14th of December).

Wessels, M.W. and Willems, P.J. (2010) 'Genetic factors in non-syndromic congenital heart malformations', *Clinical Genetics*, 78(2), pp. 103-123.

Wren, C., Birrell, G. and Hawthorne, G. (2003) 'Cardiovascular malformations in infants of diabetic mothers', *Heart*, 89(10), pp. 1217-1220.

Wu, M., Li, Y., He, X., Shao, X., Yang, F., Zhao, M., Wu, C., Zhang, C. and Zhou, L. (2013) 'Mutational and functional analysis of the BVES gene coding region in Chinese patients with non-syndromic tetralogy of Fallot', *Int J Mol Med*, 31(4), pp. 899-903.

Wu, M.C., Lee, S., Cai, T.X., Li, Y., Boehnke, M. and Lin, X.H. (2011) 'Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test', *American Journal of Human Genetics*, 89(1), pp. 82-93.

Wu, P., Teot, L., Murdoch, G., Monaghan-Nichols, A.P. and McFadden, K. (2014) 'Neuropathology of 22q11 Deletion Syndrome in an Infant', *Pediatric and Developmental Pathology*, 17(5), pp. 386-392.

Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A. and Karayiorgou, M. (2011) 'Exome sequencing supports a de novo mutational paradigm for schizophrenia', *Nat Genet*, 43(9), pp. 864-8.

Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A. and Karayiorgou, M. (2008) 'Strong association of de novo copy number mutations with sporadic schizophrenia', *Nat Genet*, 40(7), pp. 880-5.

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J.D., Romano-Adesman, A., Bjornson, R.D., Breitbart, R.E., Brown, K.K., Carriero, N.J., Cheung, Y.H., Deanfield, J., DePalma, S., Fakhro, K.A., Glessner, J., Hakonarson, H., Italia, M.J., Kaltman, J.R., Kaski, J., Kim, R., Kline, J.K., Lee, T., Leipzig, J., Lopez, A., Mane, S.M., Mitchell, L.E., Newburger, J.W., Parfenov, M., Pe'er, I., Porter, G., Roberts, A.E., Sachidanandam, R., Sanders, S.J., Seiden, H.S., State, M.W., Subramanian, S., Tikhonova, I.R., Wang, W., Warburton, D., White, P.S., Williams, I.A., Zhao, H., Seidman, J.G., Brueckner, M., Chung, W.K., Gelb, B.D., Goldmuntz, E., Seidman, C.E. and Lifton, R.P. (2013) 'De novo mutations in histone-modifying genes in congenital heart disease', *Nature*, 498(7453), pp. 220-3.

Zhang, W.J., Chan, R.J., Chen, H.Y., Yang, Z.Y., He, Y.T., Zhang, X., Luo, Y., Yin, F.Q., Moh, A., Miller, L.C., Payne, R.M., Zhang, Z.Y., Fu, X.Y. and Shou, W.N. (2009) 'Negative Regulation of Stat3 by Activating PTPN11 Mutants Contributes to the Pathogenesis of Noonan Syndrome and Juvenile Myelomonocytic Leukemia', *Journal of Biological Chemistry*, 284(33), pp. 22353-22363.

Zhao, L., Ni, S.H., Liu, X.Y., Wei, D., Yuan, F., Xu, L., Xin, L., Li, R.G., Qu, X.K., Xu, Y.J., Fang, W.Y., Yang, Y.Q. and Qiu, X.B. (2014) 'Prevalence and spectrum of Nkx2.6 mutations in patients with congenital heart disease', *Eur J Med Genet*, 57(10), pp. 579-86.

Zheng, B., Tang, T.D., Tang, N., Kudlicka, K., Ohtsubo, K., Ma, P., Marth, J.D., Farquhar, M.G. and Lehtonen, E. (2006) 'Essential role of RGS-PX1/sorting nexin 13 in mouse development and regulation of endocytosis dynamics', *Proceedings of the National Academy of Sciences of the United States of America*, 103(45), pp. 16776-16781.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. and Lander, E.S. (2014) 'Searching for missing heritability: designing rare variant association studies', *Proc Natl Acad Sci U S A*, 111(4), pp. E455-64.