Analysis and interpretation of next-generation sequencing data for the identification of genetic variants involved in cardiovascular malformation

Darren T. Houniet

For the degree of

Doctor of Philosophy

Newcastle University

Faculty of Medical Sciences

Institute of Genetic Medicine

February 2013

Abstract

Congenital cardiovascular malformation (CVM) affects 7/1000 live births. Approximately 20% of cases are caused by chromosomal and syndromic conditions. Rare Mendelian families segregating particular forms of CVM have also been described. Among the remaining 80% of non-syndromic cases, there is a familial predisposition implicating as yet unidentified genetic factors. Since the reproductive consequences to an individual of CVM are usually severe, evolutionary considerations suggest predisposing variants are likely to be rare. The overall aim of my PhD was to use next generation sequencing (NGS) methods to identify such rare, potentially disease causing variants in CVM.

First, I developed a novel approach to calculate the sensitivity and specificity of NGS data in detecting variants using publicly available population frequency data. My aim was to provide a method that would yield sound estimates of the quality of a sequencing experiment without the need for additional genotyping in the sequenced samples. I developed such a method and demonstrated that it provided comparable results to methods using microarray data as a reference. Furthermore, I evaluated different variant calling pipelines and showed that they have a large effect on sensitivity and specificity.

Following this, the NovoAlign-Samtools and BWA-Dindel pipelines were used to identify single base substitution and indel variants in three pedigrees, where predisposition to a different disease appears to segregate following an autosomal dominant mode of inheritance. I identified potentially causative variants segregating with disease in all three of the pedigrees. In the pedigrees with Dilated Cardiomyopathy and Hereditary Sclerosing Poikiloderma these variants were in plausible candidate genes.

Finally, NGS was used to identify rare, potentially disease causing indel variants in patients with sporadic, non-syndromic forms of CVM characterised by chamber hypoplasia. Two indel calling pipelines were used as a means to increase confidence in the identified indels. These two pipelines achieved the highest sensitivity calls using the method described above. In the 133 cases, evaluated for 403 candidate genes, indels were identified in 4 known causative genes for human cardiovascular disease, namely *MYL1*, *NOTCH1*, *TNNT2*, and *DSC2*.

Acknowledgements

First of all I would like to thank both my supervisors, Prof. Bernard Keavney and Dr. Mauro Santibanez-Koref, for their fantastic supervision, guidance, support and encouragement, throughout my PhD. In particular, thank you to Mauro, your perceptive nature and blunt, but seemingly always correct (!!), advice was much appreciated.

I would also like to thank everyone in the SIGM office, Prof. Heather Cordell, Dr. Ian Wilson, Yaobo Xu, Helen Griffin, Marla Endriga, Kristin Ayres, Jakris Eu-Ahsunthornwattana, Matthieu Miossec, Rebecca Darlay (or Baker), Richard Howey and Valentina Mamasoula. Your constant support has always helped. I would especially like to thank Kristin, Ian and Helen for what seemed like a constant supply of coffee, cakes and biscuits throughout my PhD. Also, thank you to Helen for your morning rants about the metro, weather and children, it got many a day started with a smile!

Thank you also to the extended members of the office and SGIM family, such as Dr. Joanna Elson and Katie Siddle. Katie remains a source of encouragement. Jo was always very encouraging and helped orchestrate many an extended lunch break where many a varied discussion occurred.

Thank you to all the members of the BHF group, Dr. Judith Goodship, Dr. Ana Topf, Dr. Thahira Rahman, Dr. Elise Glen, Dr. Darroch Hall, Mr. Rafiqul Hussein, Dr. Ruairidh Martin, Dr. Danielle Brown, Mr. Mzwandile Mbele, Ms. Helen Weatherstone. I would like to thank Ana, Thahira, Elise, and Raf for always responding to my nagging questions for data and advice, particularly Elise whom I sent a great many emails too! Mzwandile, it was great to have a fellow South African in the group! Helen, thank you for your support and endless help with anything that needed doing!

I would also like to thank both of my assessors, Heather and Dr. Simon Pearce. Your insight, support and guidance was very helpful.

Finally, thank you to my fiancé, Jen, for her constant support through my entire PhD. She put up with much moaning about my work and abilities!

Statement of contributions

Unless specified otherwise, all the work in my thesis is entirely my own. I performed all the testing and development of the method proposed in chapter 3, as well as the data analysis, variant filtering and candidate gene identification described in chapter 4 and chapter 5 of my thesis.

Dr. Mauro Santibanez-Koref dealt with the design of the statistical procedure of the method proposed in chapter 3. Dr. Matthew Hurles and Dr. Saeed Al Turki from The Wellcome Trust Sanger Institute provided 19 of the sample files used in chapter 3. They also provided the microarray data for these 19 samples.

Sequencing of 12 of the samples used in chapter 3 and all of the samples in chapter 4 was performed by Dr. Thahira Rahman, Dr. Elise Glen, and Mr. Rafiqul Hussein. Prof. Judith Goodship provided the samples and clinical information for the cases of Atrioventricular Septal Defect and Dilated Cardiomyopathy in chapter 3. Collaborators in South Africa, under Prof. Bongani Mayosi at the University of Cape Town provided samples and information for the pedigree where cases presented with Hereditary Sclerosing Poikiloderma (HSP). Collaborators in Nantes, France, under Dr. Sébastien Küry provided sequence information for the second HSP pedigree. Where possible, the identified variants were validated by Dr. Elise Glen and Dr. Thahira Rahman.

The study described in chapter 5 was part of a large, international collaboration involving researches from across Europe, called HeartRepair. All 133 samples were provided by six centres located in The Netherlands (Academic Medical Center, Amsterdam and Leids Universitair Medisch Centrum, leiden), England (The University of Newcastle, Newcastle Upon Tyne), Belgium (Katholike Universiteit, Leuven, University of Leuven, Leuven), and Germany (Max Planck Institute for Molecular Genetics and the Max Delbrück Center for Molecular Medicine in Berlin). In particular, Dr. Alex Postma (AMC, Amsterdam), Mr. Alejandro Sifrim (Katholike Universiteit, Leuven), Dr. Silke Sperling (Max Plank Institute for Molecular Genetics, Berlin), Prof. Sabine Klaasen (Max Delbrück Center, Berlin), Dr. Peter ten' Hoen (LUMC, Leiden), Dr. Thahira Rahman, Mr. Rafiqul Hussein, Dr. Ana Topf, Dr. Darroch Hall, Dr. Judith Goodship and Prof. Bernard Keavney (Newcastle University), were all involved in the sample selection, sequencing, project design, identification and validation of single base substitutions. I provided some input for the identification of the single base substitutions, however as described in chapter 5, I performed all analysis for the identification of indels.

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

1000G	1000 genomes
AD	autosomal dominant
AR	autosomal recessive
AS	association studies
AVSD	atrioventricular septal defect
BAM	binary alignment map
bp	base pair
BWA	burrows-wheeler alignment
CASAVA	consensus assessment of sequence and variation
CEU	central european
CHR	chromosome
CNV	copy number variants
CVM	cardiovascular malformation
dbSNP	single nucleotide polymorphism database
DCM	dilated cardiomyopathy
Dindel	detection of indels
DNA	deoxyribonucleic acid
EA	ebstein's anomaly
EVS	exome variant server
GA	genome analyser
GATK	genome analysis toolkit
Gb	giga base

GSNAP	genomic short-read nucleotide alignment program
GWA	genome wide association
HapMap	haplotype map
HGMD	human gene mutation database
HSP	hereditary sclerosing poikiloderma
Indel	insertion/deletion
KWE	keratolytic winter erythema
LA	linkage analysis
MHC	myosin heavy chains
MYL	myosin light chains
NCBI	national center for biotechnology information
NGS	next generation sequencing
OMIM	online mendelian inheritance in man
PAIVS	pulmonary atresia with intact ventricular septum
SAM	sequence alignment map
SIFT	sorting tolerant from intolerant
SNP	single nucleotide polymorphism
SSAHA	sequence search and alignment by hashing algorithm
UTR	untranslated region

Chapter 1. Introduction

1.1 Summary

The overall aim of my PhD was to use next generation sequencing (NGS) methods to identify rare, potentially disease causing variants involved in various diseases, particularly in Congenital Cardiovascular Malformations (CVM). I explored different aspects and uses of NGS for variant identification in three linked sub-projects which all progressed concurrently. First, I developed a novel approach to calculate the sensitivity and specificity of variant calls in NGS data using publically available SNP frequency data. I developed a simple and fast method to calculate the sensitivity and specificity of variant calls for an entire NGS data analysis pipeline. The new method generated results which were comparable to current methods requiring microarray data, without requiring such data as a reference technique. This work is reported in Chapter 3.

The knowledge gained regarding the performance of different analysis pipelines was used in the second sub-project, in which I analysed whole exome sequence data for individuals from three pedigrees using the pipelines optimised in chapter 3, namely NovoAlign-Samtools. In each of these pedigrees predisposition to a different disease appears to segregate following an autosomal dominant mode of inheritance. I identified potentially disease causing variants segregating with disease in all three of the pedigrees. In the pedigrees where cases presented with Dilated Cardiomyopathy and Hereditary Sclerosing Poikiloderma I identified potentially disease causing variants in plausible candidate genes for disease. Genes were considered as plausible candidates based on current literature and knowledge on their possible influence on the same, or similar, phenotypes. Results of these experiments, including discussion of the genes and variants identified, are presented in chapter 4.

In the third linked sub-project, I used NGS to identify rare, potentially disease causing insertion/deletion (indel) variants in patients with various congenital cardiovascular malformations. Targeted sequence data for genes believed to be involved in cardiac development was generated in 133 cases with particular subtypes of cardiovascular malformation clinically considered to represent hypoplasia of one or other of the main cardiac chambers. As indels are difficult to identify using NGS methods, I decided early on to use two indel calling pipelines as a means to increase confidence in the indels that I identified, namely BWA-Dindel and BWA-GATK. The selection of these two analysis pipelines was based on current knowledge available in the literature at the time. However, during the process, I used my assessment method, when it was fully

developed, to calculate the sensitivity and specificity values for a range of different indel calling pipelines, and found that the BWA-Dindel and BWA-GATK pipelines did indeed achieve the highest sensitivity calls. In the 133 cases, evaluated for 403 candidate genes, I discovered previously undescribed frameshifting indels that, given the strong evolutionary constraints on such indels, have a high *a priori* likelihood of being related to disease. Results of these experiments, together with discussion of the genes harbouring potentially causative indels, are presented in chapter 5

As all three of the sub-projects described in chapters 3, 4, and 5 were running concurrently from the beginning of my PhD, I considered they would be most cogently presented as standalone chapters. As such, they have all been written and developed to contain individual introduction and discussion sections. Therefore, in this general introduction I will provide a background to CVM, in particular I focus on the current knowledge of the genetic factors in such disorders. I then discuss current methods used to identify the genetic causes of disease, and why I chose to use NGS over other possible techniques. Finally, I provide a background to the different NGS platforms presently available, outline the rationale for choosing the methods used throughout the thesis, and highlight the different methods for analysing the data produced by these sequencers. In the final summary discussion of the thesis (Chapter 6), I highlight and describe what I think are the main points and conclusions of each of my data chapters, together with the limitations of the work and the prospects for further study.

1.2 Cardiovascular Malformations

1.2.1 Clinical Epidemiology

Normal cardiac development is a complex process involving various transcription factors during early development (Cresci *et al.*, 2012). The term Congenital Cardiovascular Malformation (CVM) refers to alterations in heart structure and function arising from abnormal heart development during embryogenesis (Bruneau, 2008; Rosamond *et al.*, 2008). CVM’s represent the most common birth defect, with estimates of incidence ranging from 4 to 12 per 1000 births per year (Hoffman and Kaplan, 2002). However, the exact figure is difficult to determine as many CVMs may remain subclinical and may only be detected in later life (Pierpont *et al.*, 2007; Rosamond *et al.*, 2008; Griffin *et al.*, 2009). The value also depends on the types of defect that are included in the estimation (Warnes *et al.*, 2001; Hoffman and Kaplan, 2002; Pierpont *et al.*, 2007; Bruneau, 2008; Ware and Jefferies, 2012). Figure 1.1 demonstrates the variation in incidence of cardiovascular malformations from 62 different cardiac centres from around the world (From Hoffman and Kaplan (2000)). The majority of centres report an incidence of ~7 per 1000 births per year.

The prevalence of CVM is equally difficult to calculate. A study conducted in Taiwan using health care records from the National Health Insurance program, of patients born between 2000 and 2006, estimated an overall prevalence of 13.08 per 1000 people (Wu *et al.*, 2010). Again, however, this value did depend on the types of CVM that were included. For example, if bicuspid aortic valve, which typically presents in middle age, is considered a CVM, prevalence figures will depend substantially on the age of the cohort enrolled. Atrial septal defect (ASD) is also not uncommonly diagnosed after childhood, although in this case the diagnosis is usually made during young adult life rather than middle age. In a study conducted in Yuma, Arizona, medical records from the University Medical Centre in Tucson were used to calculate the prevalence of CVM to be ~14 per 1000 people (Mayberry *et al.*, 1990). Yet another study performed in England estimated the prevalence of CVM to be 5.6 per 1000 people (Wren *et al.*, 2000). This value was calculated from records of cases born between 1985 and 1989 that were available from the diagnostic database in the regional cardiology unit at the Freeman Hospital, Newcastle upon Tyne. However, due to problems of classification and ascertainment, Pierpont et al. (2007) suggest that the true prevalence may in fact be much higher than what is calculated in these studies. Although they do not hazard a guess as to what this may be.

The two most common forms of CVM are bicuspid aortic valve defects, occurring in 1% - 2% of the population, followed by septation defects, see figure 1.2 (Bruneau, 2008; Silversides *et al.*, 2010), although bicuspid aortic valve defects are often removed from estimates of prevalence. As can be seen from figure 1.2, there are many different types of CVM and various morphological variations (Hoffman and Kaplan, 2002; Rosamond *et al.*, 2008). Examples include ventricular septal defects, atrial septal defects, atrioventricular septal defects and pulmonary stenosis (Hoffman and Kaplan, 2002). All of these different variations and categories make accurate estimates of incidence and prevalence difficult to ascertain. Since the focus of my work is chiefly on data analysis in NGS, I do not discuss the clinical aspects of the different types of CVM in detail here.

Mortality rates due to CVM vary depending on the defect and its severity (Bruneau, 2008). For infants born between 1979 and 1997 in the USA, 24% of CVM associated deaths arose from patients suffering hypoplastic left heart syndrome, and 5.6% of deaths were from transposition of great artery sufferers (Boneva *et al.*, 2001). However, despite this variation, up to 85% of CVM patients now survive to adulthood (Warnes *et al.*, 2001). This is due to dramatic improvements in surgery that have increased survival rates, and now the number of adults with the disease exceeds the number of children (Burn *et al.*, 1998; Marelli *et al.*, 2007; Pierpont *et al.*, 2007; Bruneau, 2008; Rosamond *et al.*, 2008). However, despite advancements in the field, patients can suffer from secondary complications later in life, in particular neurological disorders and arrhythmias (Bruneau, 2008).

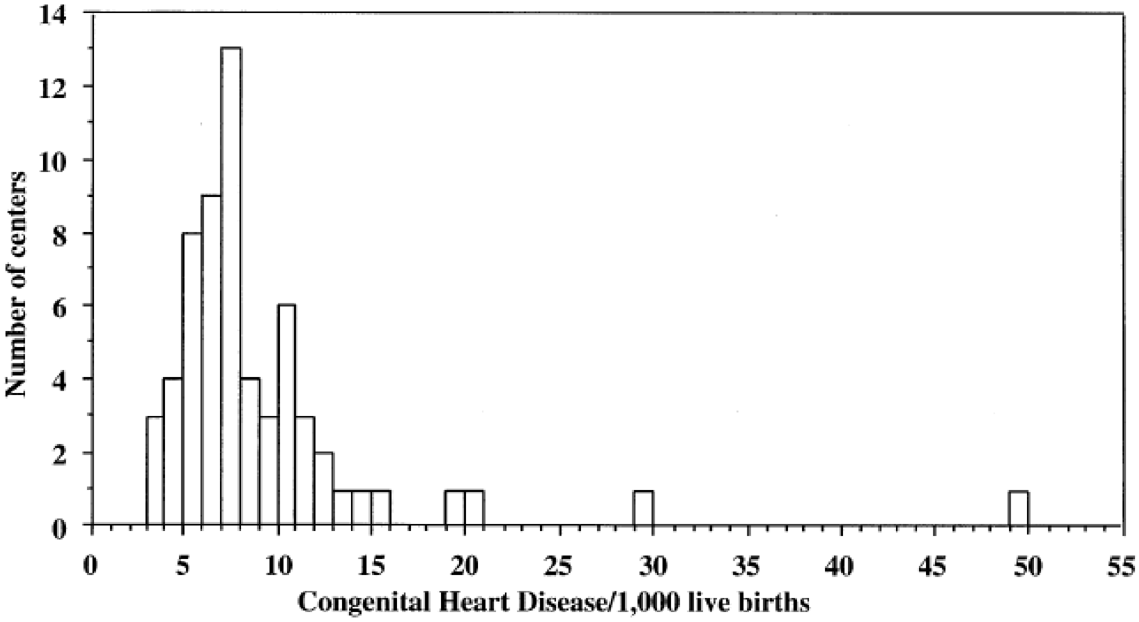


Figure 1.1. Incidence of congenital heart disease per 1000 live births. This table is from Hoffman and Kaplan (2002) who collated data from 62 cardiac centres.

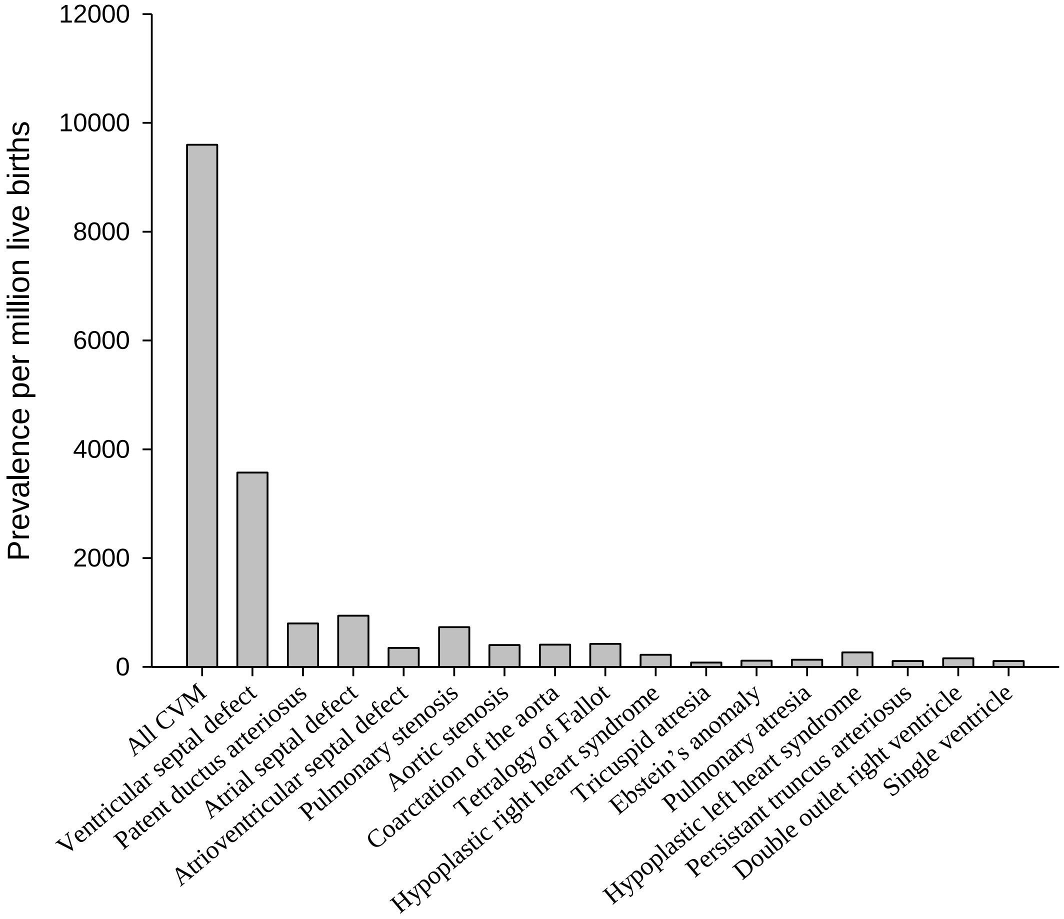


Figure 1.2. Prevalence per million live births of different types of CVM. Values are on the log scale. The graph is based on original data from Hoffman and Kaplan 2002. The "All CVM" column excludes cases of bicuspid aortic valve.

1.2.2 Risk factors

The known risk factors of CVM only account for ~20% cases (Jenkins *et al.,* 2007; Pierpont *et al.,* 2007; Griffin *et al.,* 2009). Most of what is known is based largely on studies involving Mendelian syndromes and rare, familial, non-syndromic forms of CVM. The origin of most CVM (The remaining 80%) cases is unknown with various studies suggesting both environmental and genetic causes (Jenkins *et al.*, 2007; Pierpont *et al.*, 2007; Bruneau, 2008; Ware and Jefferies, 2012). Some of the known risk factors will be discussed in more detail below.

1.2.3 Environmental risk factors

There is much evidence surrounding the possible influence of external risk factors on CVM and an understanding of these factors is important because it could lead to the possible prevention of a small number of cases (Jenkins *et al.*, 2007). One example is exposure to angiotensin-converting-enzyme inhibitors early on in pregnancy, which has been demonstrated to result in an increased CVM risk. In one such study, 209 children that were exposed only to ACE inhibitors during the first trimester were identified, as well as 202 children exposed to other antihypertensive medications in the first trimester, and 29096 children which were not exposed to any antihypertensive drugs during any time of pregnancy (Cooper *et al.*, 2006). Information from this study was obtained from the Tennessee Medicaid database of children born between 1985 and 2000. The study found an increased risk of major congenital malformations, which included CVM, in children exposed to ACE inhibitors, as compared to the group of children not exposed to any antihypertensive drugs.

Another example of a risk factor for CVM is phenylketonuria in the mother. In one study, a group of 416 children from 412 maternal phenylketonuria pregnancies were compared to 100 children from 99 control pregnancies (Levy *et al.*, 2001). Of these, CVM was identified in 34 (14%) of the children born from the phenylketonuric mothers, in comparison to only 1% from the 100 children of the control group. In addition, an article by Jenkins *et al.* (2007) reviewed the available literature on prenatal and parental conditions and exposures and associated risk of CVM up until 2006. This article highlighted many of the additional risk factors for CVM, such as maternal rubella, diabetes, and alcohol.

Maternal rubella has been associated with various cardiac defects, such as pulmonary valve abnormalities, peripheral pulmonary stenosis, and ventricular septal defects (Jenkins *et al.*, 2007). In particular, one such study reviewed literature describing congenital heart defects following maternal rubella between 1941 and 1961 (Way, 1967). Congenital defects were found in 4 – 58% of the cases. Patent ductus arteriosuis was the most common (58% of cases) cardiovascular defect seen.

Maternal pregestational diabetes has also been shown to increase the risk of CVM. Specific CVMs associated with maternal pregestational diabetes include laterality and looping defects, transposition of the great arteries, and ventricular septal defects (Becerra *et al.*, 1990; Jenkins *et al.*, 2007). It has been found that children born from mothers with gestational diabetes mellitus and who required insulin during the third trimester of pregnancy, were 20 fold more likely to be born with a major cardiovascular malformation when compared to children of non-diabetic mothers (Becerra *et al.*, 1990). The study used information on 4929 still born and live babies born between 1968 and 1980, obtained from the Metropolitan Atlanta Congenital Defects Program. The study also included 3 029 healthy babies which were matched based on ethnicity, period of birth and hospital of birth.

Alcohol has been shown to be responsible for a wide range of teratogenic affects during pregnancy, including cardiac defects (Jenkins *et al.*, 2007). Although the risks do appear to be related to the amount of alcohol consumed during pregnancy. A case-control study examining 4705 case mothers and 4329 control mothers found that although sporadic, low doses of alcohol may increase the risk of congenital defects, these risks do increase with increasing alcohol exposure. Despite various environmental factors having been shown to influence CVM risk, the focus of my PhD is on the genetic risk factors.

1.2.4 Genetic Epidemiology

CVM can be caused by a range of genetic variants, including rare, highly deleterious variants resulting in Mendelian forms, and common variants with a weak affect that can modulate the risk of complex disease (Manning *et al.*, 2005). Below I will provide a brief introduction to some of the Mendelian forms of CVM and their associated genetic risks, followed by an introduction to the sporadic, non-syndromic forms of CVM. As mentioned earlier, most knowledge on the genetic factors influencing CVM has been obtained from studies on Mendelian, syndromic and rare familial forms of non-syndromic CVM.

1.2.5 Mendelian forms

The recurrence of some CVM in families provides evidence for a genetic influence on many of these defects (Bruneau, 2008; Faita *et al.*, 2012). In fact, familial recurrence has been shown for both non-syndromic and syndromic forms of CVM (Wolf and Basson, 2010). With regards to inherited CVM, both monogenic (Single gene inheritance) and complex, polygenic (Multiple gene inheritance) forms occur (Faita *et* *al.*, 2012). Information on the frequency of recurrence within these inherited CVMs could help to improve understanding of the disease as well as provide the information necessary to help affected parents wanting to have children of their own (Burn *et al.*, 1998). Some of the inherited syndromes displaying CVM include Alagille syndrome, Noonan syndrome and Holt-Oram syndrome. These conditions as well as the genetic contributions to each will be discussed in more detail below.

Alagille syndrome is an autosomal dominant disorder presenting with various abnormalities in the liver, heart, skeleton and eyes (Pierpont *et al.*, 2007). In a study involving 222 cases, 94% displayed some form of cardiovascular abnormality (McElhinney *et al.*, 2002). The most common cardiovascular abnormalities in patients with Alagille syndrome include peripheral pulmonary hypoplasia, tetralogy of Fallot, and pulmonary valve stenosis (Pierpont *et al.*, 2007). Alagille syndrome can result from deletions of chromosome 20p12 or from mutations in the *JAG1* gene (McElhinney *et al.*, 2002; Pierpont *et al.*, 2007). As an example, in one such study the *JAG1* gene was analysed in four families where 10 members suffered from Alagille syndrome (Li *et al.*, 1997). In this study RT-PCR products were screened for mutations by heteroduplex mobility analysis. Four coding variants were identified in *JAG1* which segregated with disease and were identified in all four families, but not in 100 controls.

Another example is Noonan syndrome, in which 50 - 90% of patients are affected by cardiac disorders (Manning *et al.*, 2005; Pierpont *et al.*, 2007; Ware and Jefferies, 2012). This is an autosomal dominant disorder characterised by typical facies, pterygium colli, short stature and CVM (Marino and Digilio, 2000). Three disease genes in the RAS-MAP Kinase signalling pathway have been shown to influence Noonan syndrome, namely *PTPN11*, *SOS1*, and *KRAS* (Pierpont *et al.*, 2007). For example, a genome-wide linkage analysis in a large Dutch pedigree was able to identify the distal part of chromosome 12q as being linked to disease (Jamieson *et al.*, 1994).

Holt-Oram syndrome is an autosomal dominant disorder which is characterised by CVMs in patients with upper limb deformities (Pierpont *et al.*, 2007). In a study involving 55 Holt-Oram cases, and their parents, 95% of the cases displayed some form of cardiac defect (Newbury-Ecob *et al.*, 1996). The various cardiac disorders included atrial septal defects and ventricular septal defects. Studies have been able to link variants in the *TBX5* gene to Holt-Oram syndrome (Pierpont *et al.*, 2007). For example, a genome wide linkage analysis was used to identify mutations in the *TBX5* gene as responsible for Holt-Oram syndrome in two large families (Basson *et al.*, 1994).

As well as occurring as part of recognised syndromes, CVM's can occur as isolated (or non-syndromic) inherited defects (Richards and Garg, 2010). For example, genetic variants within the *NKX2-5* gene have been linked with non-syndromic CVM (Schott *et al.*, 1998). In this study, four families displaying high incidences of CVM were analysed. In all four families the disease displayed a pattern of inheritance consistent with autosomal dominant transmission. A genome wide linkage analysis identified a region on chromosome 5, where the *NKX2-5* gene is located. The exons of all genes in the linkage region were sequenced in all the affected individuals across all four families, with the *NKX2-5* gene being identified as the only gene to have a shared variant in all affected family members. There are also many other cases describing genetic variants responsible for non-syndromic CVM. For instance, mutations in the *GATA4* gene have been linked to cases of isolated CVM (Garg *et al.*, 2003), and mutations in the *MYH6* gene have also been linked to dominantly inherited atrial septal defects (Ching *et al.*, 2005).

1.2.6 Sporadic, non-syndromic forms

The genetic causes of sporadic, non syndromic forms of CVM are far more difficult to identify (The reasons are discussed in more detail below). However there have been some successes. For example, d*e novo* copy number variants (CNVs) have been associated with tetralogy of Fallot (Greenway *et al.*, 2009; Soemedi *et al.*, 2012). In the study by Soemedi *et al.* (2012), the frequency of CNVs in 2 256 CVM cases was compared with 841 controls. They were able to identify significant differences in the deletion burden between the cases and the controls. As well as CNVs, other deletion/duplication events have also been shown to share an association with CVM (Bruneau, 2008; Ware and Jefferies, 2012), and chromosomal aberrations appear to occur fairly frequently in some CVM cases (Pierpont *et al.*, 2007). This has been demonstrated in a study examining infants born with CVM between 1981 and 1986, where ~13% of them displayed chromosomal abnormalities (Ferencz *et al.*, 1989).

1.2.7 Genetic study approaches

Linkage Analysis (LA) and Association Studies (AS) have commonly been used to try and identify the genetic causes of disease (Bailey-Wilson and Wilson, 2011). More recently, however, next generation sequencing approaches, sometimes in conjunction with linkage or association methods, have been employed to identify the genetic causes of various disorders. However, each of these methods have their own particular strengths and weaknesses and perform best depending on the disease. Faita *et al.* (2012) provide a diagram (Figure 1.3) describing the genetic contribution to monogenic and multigenic CVMs and the study approaches which can be adopted to identify these. In the case of complex diseases, many genetic variants occur at a higher population frequency and can be identified using a genome wide association approach. In the case of Mendelian, monogenic disorders the diseases are caused by rare mutations in specific genes, for which a linkage study can be used. NGS may provide a means of identifying the genetic causes of both monogenic and complex diseases.

LA has been used highly successfully for identifying many thousands of disease causing variants in a range of Mendelian diseases, including various cardiac disorders (Bailey-Wilson and Wilson, 2011; Parvez and Darbar, 2011). For example, LA allowed for the identification of variants in chromosomes 18q22, 13q34 and 5q21 which are linked to bicuspid aortic valve disease (Hinton *et al.*, 2009), and LA was also used to identify a missense mutation in the *JAG1* gene in families with Alagille syndrome (Eldadah *et al.*, 2001).

LA often precedes approaches such as exome sequencing as it allows for all of the variants outside of linkage peaks to be excluded, hopefully reducing the number of candidate variants to a more manageable number (Bailey-Wilson and Wilson, 2011; Smith *et al.*, 2011). These linkage peaks identify regions of identity-by-descent matching a particular genetic model (Smith *et al.*, 2011). LA in combination with whole exome sequencing has been employed to identify many potentially causative variants in autosomal dominant and autosomal recessive disorders (Smith *et al.*, 2011), for example in the case of the dominantly inherited Amyotrophic Lateral Sclerosis, ALS (Johnson *et al.*, 2010a). That study describes a multigenerational family in which four members presented with ALS. Exome sequencing was performed on two of the affected family members in which 88 variants were identified. Of the 88 identified variants, 33 were validated using Sanger sequencing. Of these 33, only four were present in the *VCP* gene with a LOD (Logarithm of the odds) score above zero. All four variants were predicted by SIFT to be disease causing, and none were present in a group of 200 healthy controls.

LA and exome sequencing approaches have also been successfully employed to identify the causative variants in many recessive disorders, as in the case of autoimmune lymphoproliferative syndrome, ALPS (Bolze *et al.*, 2010). That study describes a pedigree with multiple affected individuals and in which the disease appears to display an autosomal recessive inheritance pattern. Three patients were genotyped for use in the LA, and whole exome sequencing was performed on one of these individuals. LA identified three regions of homozygosity shared by the patients. All three sites were heterozygous in healthy members of the family. Of the 23146 variants that were identified, only 81 were located in the candidate regions. After further filtering, only 1 non-synonymous variant remained which was not seen in a group of 282 healthy controls.

Despite the various successes, LA only has the power to detect alleles with large affect sizes, which are rare in populations, and which have high penetrance (Bailey-Wilson and Wilson, 2011). For complex diseases, an alternative approach would be to use a Genome Wide Association (GWA) study (Zhu and Xiong, 2012). Due to the availability, and decreased costs of genotyping using microarrays, GWA studies have recently become a commonly used method in complex disease research (Paynter *et al.*, 2010; Bailey-Wilson and Wilson, 2011; Zhu and Xiong, 2012).

GWA studies are used to identify associations between common alleles and disease phenotypes (Bailey-Wilson and Wilson, 2011) and have proved useful in the search for the genetic causes of various different types of CVM (Arking and Chakravarti, 2009; Paynter *et al.*, 2010). For example, in the study by Schott *et al.* (1998), a GWA identified *NKX2-5* as a possible gene causing various CVM's. Also, a GWA identified a region on chromosome 12q24 as being associated with tetralogy of Fallot (Cordell *et al.*, 2013 (In press)). Other successful studies include a study in which 7 SNPs were found to be associated with cardiovascular disease (Smith *et al.*, 2010), and another study identified 3 genes associated with coronary artery disease (Feng and Zhu, 2010).

However, GWA studies are based on the common disease-common variant model, assuming that common variants are likely to be important factors in common disease (Reich and Lander, 2001; Juran and Lazaridis, 2011). Supposing that individual rare variants only have a small affect on common disease, and also that they have very low population frequencies, present methods for testing for associations have limited power (Zhu and Xiong, 2012). Therefore, GWA studies are better powered for common SNPs displaying a frequency of ~5% and higher (Hindorff *et al.*, 2009; Manolio *et al.*, 2009; Cirulli and Goldstein, 2010). However, some are now including variants with a frequency as low as ~1% (Manolio *et al.*, 2009; Cirulli and Goldstein, 2010), but association testing using alleles with frequencies of <0.5% still has very low power, unless the effect size of the allele is very large (Manolio *et al.*, 2009). Traditional GWA studies also assume a degree of sample independence and are therefore unable to assess correlated family data (Zhu and Xiong, 2012). Although more modern association methods may allow for case-control tests within pedigrees, they still require the allele frequencies to be high (Zhu and Xiong, 2012). Additionally, Manolio *et al*. (2009) highlight the importance of sample size in GWA studies, stating that much larger sample sizes are required to detect associations using very rare alleles.

Despite the many success stories using GWA and LA methods, the identified variants only appear to explain a very small proportion of the heritability of disease (Zuk *et al.*, 2012). Here I define heritability as the proportion of phenotypic differences in a population that are explained by genetic factors (Manolio *et al.*, 2009). For example, in atrial fibrillation, both GWA and LA have been used to successfully identify causative genetic variants, but these genes only explain less than 10% of the genetic heritability of the disease (Parvez and Darbar, 2011). This phenomenon has been termed “missing” heritability, and the current understanding is that this missing heritability lies in additional variants which have not been discovered yet (Maher, 2008; Makowsky *et al.*, 2011; Zuk *et al.*, 2012).

Manolio *et al.* (2009) provides a diagrammatic representation of the possible relationship between allele frequency and effect size, figure 1.4. The diagram describes the difficulty of identifying very rare variants (allele frequency <0.001) with low effect sizes. Other explanations include the difficulty in identifying and genotyping certain types of variants such as CNVs, a low power to detect gene-gene interactions, and the inadequate identification of possible, shared environmental factors (Maher, 2008; Manolio *et al.*, 2009).

Both LA and GWA studies have demonstrated that the “missing heritability” cannot be explained by rare, large effect alleles, or by common moderate effect alleles (Manolio *et al.*, 2009). However, it is important to be able to correctly identify risk alleles, and the source of this missing heritability (Paynter *et al.*, 2010), as it may lead to a better understanding and treatment of the disease (Manolio *et al.*, 2009). Some authors have claimed that sequencing could be used to investigate the presence of causative rare variants further, and as a means of identifying the sources of this missing heritability (Manolio *et al.*, 2009).

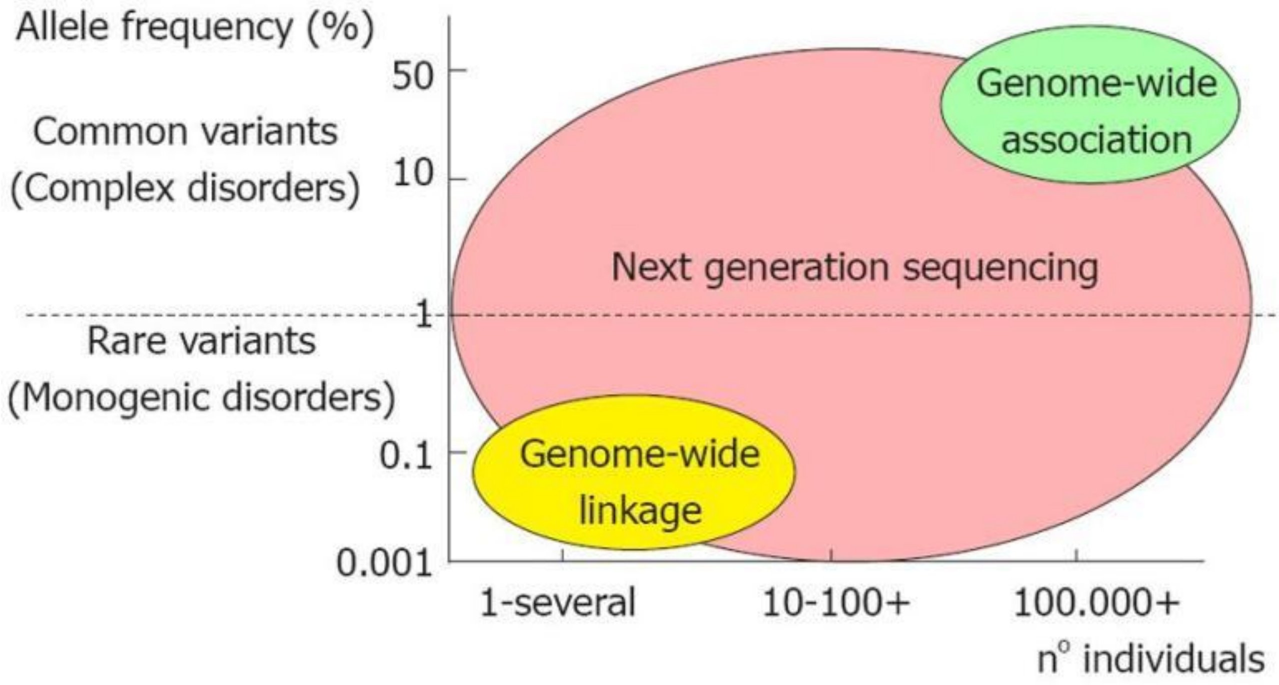


Figure 1.3. Genetic contribution to monogenic and complex CVM (From Faita et al. (2012)).

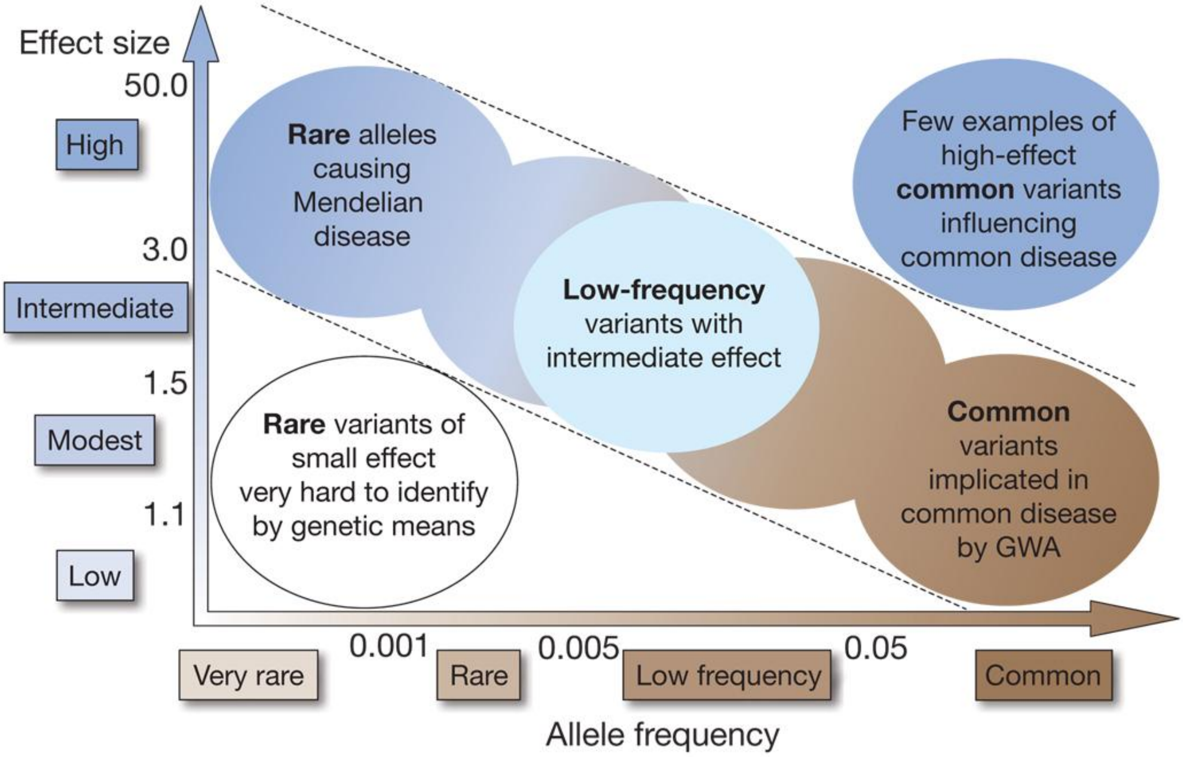

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 1.4. Identifying genetic variants based on risk allele frequency and strength of genetic effect. The figure is from Manolio et al. (2009).

1.3 NGS in Mendelian Diseases

NGS approaches have been particularly successful in identifying the causes of various Mendelian diseases displaying a recessive mode of inheritance (Lalonde *et al.*, 2010; Ng *et al.*, 2010b; Bamshad *et al.*, 2011; De Keulenaer *et al.*, 2012; Pyle *et al.*, 2012). This is largely because homozygous, disease causing variants in populations occur rarely (Ng *et al.*, 2010b; Ng *et al.*, 2010c; Stitziel *et al.*, 2011). For example, in the case of spastic ataxia of Charlevoix-Saguenay (Pyle *et al.*, 2012). In this study, the authors were able to identify the *SACS* gene as the likely cause of disease by using the exomes of only two cases. Many of the other studies quoted here are described in more detail in chapter 4.

In chapter 4 I used NGS methods to identify rare variants in diseases appearing to show a dominant mode of inheritance. Although, more difficult than causal variant discovery in recessive conditions (See chapter 4), various studies have shown a degree of success in identifying the causative variants responsible for dominant disease (Johnson *et al.*, 2010a; Ng *et al.*, 2010a; Dickinson *et al.*, 2011; Pfeffer *et al.*, 2012). As with the recessive studies listed above, these studies are discussed in more detail in chapter 4.

1.4 NGS in Complex Diseases

As well as identifying potentially disease causing variants in Mendelian conditions, I also used NGS methods to identify potentially disease causing variants in sporadic forms of CVM (Chapter 5). Of particular interest in this regard was an empirical study of exome sequence data from 438 individuals, comprising 184 individuals from the International HIV Controllers Study and 254 control individuals (Kiezun *et al.*, 2012). In this study several of the important aspects of exome sequencing for disease identification in non-Mendelian diseases were discussed, in particular the problem of obtaining adequate sample sizes. Using simulated data, the authors expect that over 10000 exomes would be required to achieve sufficient statistical power to detect associations of rare variations with complex traits. Also, the many GWA studies which have been carried out to date, demonstrate that common variants underlying complex traits are not necessarily located in exonic gene regions, but spread across many more regions of the genome (Day-Williams and Zeggini, 2011). However, protein coding genes do provide well defined and easily interpretable targets (Kiezun *et al.*, 2012), particularly in genes known to be involved in disease, such as the targeted approach I used in chapter 5.

Despite these difficulties, there have been some successes, such as in Type I diabetes (Nejentsev *et al.*, 2009). In this study exons and splice sites of 10 candidate genes in 480 cases and 480 controls were re-sequenced using a 454 GS-FLX sequencer (Described below). A total of 212 point mutations were identified using the Staden package (http://staden.sourceforge.net/). Of these, 179 were categorised as rare (Minor allele frequency <3%), of which 156 had been previously unreported. The authors tested for association using all 212 variants by comparing cases and controls, and were able to confirm the previously known associations of common SNPs with type 1 diabetes. However, they also identified associations with two rarer SNPs, rs35667974 and rs35337543, in the *IFIH1* gene. Both variants were predicted to alter the expression and structure of the gene.

1.5 NGS platforms

1.5.1 Sanger sequencing

Since its inception, the Sanger sequencing method has been used in thousands of ground breaking studies, such as in the sequencing of the first human consensus sequence (Lander *et al.*, 2001; Harismendy *et al.*, 2009). Sometimes called first generation sequencing, this method employs dideoxy and arabinonucleoside analogues which act as chain terminating inhibitors of DNA polymerase (Sanger *et al.*, 1977). This process can generate reads of up to 900 base pairs in length (Zhang *et al.*, 2011; Liu *et al.*, 2012) and is frequently used in validation studies.

Although still considered by many as the "gold standard" sequencing approach, in comparison to more modern approaches it is expensive and has a considerably lower output, see table 1.1 (Harismendy *et al.*, 2009; Audo *et al.*, 2012). The recent advancements in NGS technologies have made it possible to sequence entire genomes and exomes in a relatively short space of time. It is now possible to examine variation in multiple genomic segments of several samples (Choi *et al.*, 2009), which could greatly improve understanding of the genetics behind complex diseases (Choi *et al.*, 2009; Ng *et al.*, 2009), especially in the identification of rare variants (Choi *et al.*, 2009; Cirulli and Goldstein, 2010; Parvez and Darbar, 2011). These methods are able to rapidly generate millions of sequence reads per patient and provide a fast, cost effective means to identify potentially disease causing variants in CVM cases (Mardis, 2008; Arking and Chakravarti, 2009).

1.5.2 Roche 454 GS-FLX system

The 454 GS-FLX system was the first commercially available NGS platform. It utilised an emulsion PCR amplification stage followed by a sequencing-by-synthesis technique whereby reagents flowing across a slide allow for simultaneous nucleotide extension reactions, with each base incorporation emitting a light signal captured by a camera (Margulies *et al.*, 2005; Mardis, 2008; Pareek *et al.*, 2011; Zhang *et al.*, 2011).

Initially the genome is sheared to produce random libraries of DNA fragments. Adaptors are added to each fragment, and the fragments are then captured on beads. These are then clonally amplified in an emulsion PCR step, and the resultant DNA strands are denatured and deposited into wells of a fibre-optic slide. The slide is deposited in a flow chamber where sequencing-by-synthesis occurs. Reagents simultaneously flow across the flow chamber allowing for extension reactions where nucleotide incorporation results in the release of a light signal which is detected by a camera. The intensity of the light signal depicts the number of nucleotides which were added at each reaction (Margulies *et al.*, 2005). The most recent 454 GS-FLX system is able to generate reads of up to 700bp long (Liu *et al.*, 2012). Although its throughput is lower than for the other NGS systems, it is much faster than both the SOLiD and HiSeq sequencers and able to generate much longer reads (Liu *et al.*, 2012). The 454 GS-FLX system is often used for *de novo* sequence assembly, cancer and mutation detection applications (Liu *et al.*, 2012).

Roche has also released a bench top sequencer called the GS Junior, in which the library preparation and data processing has been simplified and the system can now generate up to 14Gb per run (Liu *et al.*, 2012). This sequencer is able to generate up to 100 000 reads with an average read length of ~400bp's (http://www.gsjunior.com/instrument-workflow.php).

1.5.3 Applied Biosystems SOLiD

The Applied Biosystems SOLiD Sequencer employs an emulsion PCR amplification step, before DNA sequencing using fluorescently labelled dinucleotides that are added by ligation (Mardis, 2008). This system has many applications, including whole genome resequencing, targeted resequencing, transcriptome research (including gene expression profiling and small RNA analysis), and epigenomic research (Liu *et al.*, 2012).

Initially, the genomic DNA is fragmented and oligo adaptors ligated to the ends of the DNA fragments. The adaptor sequences are then hybridised to magnetic beads containing complementary oligos and the sequences are amplified via emulsion PCR. The beads are then attached to the surface of a glass slide and placed in the sequencer. During sequencing, a universal primer, complementary to the adapters, is annealed to the library fragments. A set of 8mer oligonucleotides and DNA ligase is then added. If an oligonucleotide hybridises to the DNA fragment sequence next to the 3' end of the primer, the DNA ligase seals the phosphate backbone. After ligation, the 8mer oligonucleotide is identified by a fluorescent label on the fifth or second base position.

Finally the 6th through 8th bases of the oligonucleotide sequence are removed by cleavage, allowing for another round of ligation. The DNA fragment sequence is therefore identified in steps of five nucleotide intervals (Mardis, 2008). The ABI SOLiD 5500xl system can now generate reads of up to 85bp in length and has a total output of 30Gb per run (Liu *et al.,* 2012).

1.5.4 Illumina Genome Analyser

The Illumina Genome Analyser (Illumina GA) was the second commercially available NGS platform and utilised cluster amplification and a sequencing-by-synthesis technique using reversible fluorescently labelled chain terminators (Mardis, 2008; Zhang *et al.,* 2011).

Initially the single stranded genomic DNA is fragmented and adaptor oligonucleotides ligated to the individual fragments. The fragments are then added to the surface of a glass flowcell comprising 8 lanes with complementary, covalently attached oligos. The fragments can then hybridise to the flow cell oligos and then undergo PCR amplification in a cluster. The flowcell is transferred to the sequencer and supplied with polymerase and fluorescently labelled nucleotides. The fluorescent label identifies the base. The 3' end of each base is inactivated insuring the addition of only one base per cycle. Each cycle is followed by an imaging step to identify the particular base that was incorporated and the fluorescent group and the 3' block are removed (Mardis, 2008). The original Illumina GA was able to output up to 95Gb of sequence data, comprising reads of up to 150bp long (http://www.illumina.com/).

Then in early 2010, Illumina launched the HiSeq2000 which due to various improvements in polymerase, buffer, flowcell, and software, is now able to output up to 600Gb per run (Liu *et al.,* 2012). Like its predecessor, the HiSeq2000 employs a sequencing-by-synthesis approach, but at a 2 – 5 fold higher rate of data acquisition by using a four camera system able to detect the intensities of all four bases simultaneously (Minoche *et al.,* 2011). Additionally, in comparison to the Roche 454 GS-FLX system and the Applied Biosystems SOLiD systems, the HiSeq2000 is also the cheapest on the market, with a sequencing cost of only $0.02/million bases (Liu *et al.,* 2012). The HiSeq2000 can be used for a range of different applications; in particular it is frequently used in targeted reqequencing and mutation discovery studies.

Illumina has also released its own compact, bench top sequencer called the MiSeq. This sequencer still makes use of Illumina's reversible terminator-based sequencing by synthesis technique and is able to generate reads of up to 150bp in length and output up to 1.5Gb of data (Liu *et al.,* 2012). This lower throughput, fast turnaround sequencer has been largely aimed at small laboratories and for use in clinical diagnostics (Quail *et al.,* 2012).

1.5.5 Life technologies Ion Torrent PGM

Clonal amplification of the DNA fragments is achieved via an emulsion PCR step on the surface of 3-micron diameter beads, called Ion Sphere Particles (Quail *et al.*, 2012). These beads are then loaded into proton-sensing wells on a silicone wafer, after which each of the four bases is introduced sequentially (Zhang *et al.*, 2011; Quail *et al.*, 2012). During the sequencing process, when a nucleotide is incorporated onto the growing DNA strand, a hydrogen ion is released which is detected by an ion sensor and converted into a digital output (Zhang *et al.*, 2011). The direct connection between chemical and digital information improves speed, simplicity and output (Pareek *et al.*, 2011).

However, the system does carry with it some disadvantages, as listed by Niedringhaus et al. (2011). The author's state that the need for the reaction wells to be cleared between each reaction step can lead to an accumulation of errors, and that the system has difficulties sequencing highly repetitive or homopolymer regions.

1.5.6 Third generation sequencing

As well as the next generation platforms described above, methods utilising single DNA molecule sequencing are currently being made available (Zhang *et al.*, 2011). These methods do not require amplification and simply read through DNA templates in real time, making them potentially more accurate than the NGS platforms (Pareek *et al.*, 2011; Zhang *et al.*, 2011; Liu *et al.*, 2012).

For example, the Pacific Biosciences RS which relies on a process termed single molecule real time sequencing that employs a sequencing-by-synthesis method which uses fluorescently labelled nucleotides and DNA templates attached to the bottom of zero-mode waveguide wells (Pareek *et al.*, 2011; Quail *et al.*, 2012). During this process the DNA templates are attached to the bottom of the 50nm wide zero-mode waveguide wells (Quail *et al.*, 2012). DNA synthesis is carried out by DNA polymerase in the presence of y-phosphate fluorescently labelled nucleotides (Pareek *et al.*, 2011; Quail *et al.*, 2012). With each new base incorporation, the flourophores attached to the nucleotides are excited, generating a pulse of fluorescence which is detected in real time (Quail *et al.*, 2012).

Some of the advantages of this method of sequencing are listed by Liu *et al.* (2012). These advantages include a decreased sample preparation time of only 4 - 6 hours, a PCR step is not required which reduces errors caused by PCR, the turnover rate is very fast (An entire run can be completed in a day) and finally, the average read length is ~1300bp. Despite these apparent advantages, the system suffers from inefficient loading of the DNA polymerase into the zero-mode waveguide wells and subsequent degradation of the polymerase in these wells, a low accuracy of between 81 – 83%, and a high cost per base (Niedringhaus *et al.*, 2011).

1.5.7 Comparison of different sequencing platforms

Harismendy *et al.* (2009) compared the base calls generated using Sanger sequencing to those generated using the 454 GS-FLX, Illumina GA and the ABI SOLiD platforms. They identified heterozygous and homozygous variants in 258 879 base pairs using all four methods and found 20 loci that the three NGS technologies were concordant with, but discordant with the Sanger calls. Eight of these 20 calls were base calling errors in the original samples, while 9 of the remaining 12 discrepancies were found to be incorrect in the Sanger sequences. Sanger sequencing had a 0.9% false positive call rate and a 3.1% false negative call rate, whereas the 454 GS-FLX, Illumina Genome Analyser and ABI SOLiD displayed false positive rates of 2.5%, 6.3% and 7.8% respectively, and false negative rates of 3.1%, 0%, and 0.9% respectively.

Liu *et al.* (2012) also provided a comparison between the Sanger sequencing platform and three NGS platforms, namely the 454 GS-FLX system, the Illumina HiSeq2000 system, and the SOLiD 5500xl system (Table 1.1). This table and its information was extracted from Liu *et al.* (2012). As can be seen, the most significant differences between the systems are in terms of read length, data output and cost. By far, the most expensive, with the lowest throughput, is Sanger sequencing. However, it does generate very long reads and displays a very high accuracy.

A further study by Niedringhaus *et al.* (2011) compared the Sanger, 454 GS-FLX, HiSeq2000, and SOLiD sequencing platforms. They too found that although the Sanger sequencing method was more expensive and had a lower throughput than the next generation platforms, it could generate longer, high quality reads that also displayed good quality in repeat and hompolymer regions. They found that the sample preparation step was complicated for the 454 GS-FLX system and that it generated low quality reads in repetitive and homopolymer regions. However, the 454 GS-FLX sequencer was able to generate longer reads than any of the other next generation platforms. In comparison, the SOLiD sequencer produced relatively short reads amongst, but had a very high throughput and very low reagent costs.

Quail *et al.* (2012) provides a comparison of three next generation bench-top sequencers, namely the Illumina Miseq, the Ion Torrent, and the PacBio RS (Table 1.2). Although the Illumina MiSeq has the lowest costs and highest output, it also has the shortest read length and longest run time. However, the MiSeq also has the lowest error rate at 0.8%, with the PacBio RS sequencer having a much higher error rate at 12.86%.

Due to its very high output, very low per base cost, and its versatility I chose to use the Illumina Genome Analyser and Illumina HiSeq2000 as the sequencers of choice in chapters 4 and 5 of my thesis.

Sequencer	454 GS FLX	HiSeq2000	SOLiD 5500xl	Sanger 3730xl
Read length	700 bp	50SE, 50PE, 100PE	85bp	900 bp
Accuracy	99.90%	98%, (100PE)	99.99%	100.00%
Output data/run	0.7 Gb	600 Gb	30 Gb	84 Kb
Time/run	24 Hours	~8 Days	7 Days for SE or 14 Days for PE	20 Mins - 3 Hours
Cost/million bases	$10	$0.02	$0.13	$2400
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

Table 1.1. Comparison of the 454 GS FLX, HiSeq2000, SOLiD 5500xl and Sanger 3730xl sequencing systems. Table and information has been extracted from Liu *et al.* (2012). SE=Single-end and PE=Paired-end.

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS
Instrument Cost	$128K	$80K	$695K
Sequence yield per run	1.5-2Gb	20-50Mb on 314 chip, 100-200Mb on 316 chip, 1Gb on 318 chip	100Mb
Sequencing cost per Gb	$502	$1000 (318 chip)	$2000
Run Time	27 hours	2 hours	2 hours
Observed Raw Error Rate (%)	0.8	1.71	12.86
Read length	up to 150 bases	~200 bases	Average 1500 bases

Table 1.2. Comparison of three next generation, bench-top sequencing platforms. The table and information has been adapted from Quail *et al.* (2005).

1.6 NGS Data Analysis

Following sequencing, a series of programmes are employed to convert the sequencer output into a nucleotide sequence, and ultimately to identify variants, from the sequencer output files. A base caller must first be used to convert the output into sequence reads and to assign the correct base identity to each sequence (Nielsen *et al.*, 2011). A sequence aligner is then used to align the sequence reads to a reference sequence, and remove miscellaneous sequences not matching the reference sequence (Day-Williams and Zeggini, 2011). Finally a variant caller can be used in order to identify variants (Which I define as deviations from the reference sequence) from these aligned reads (McKenna *et al.*, 2010; Koboldt *et al.*, 2012).

There are a whole host of programmes available to perform sequence alignment (Li and Durbin, 2009; Wu and Nacu, 2010; Langmead and Salzberg, 2012) and variant calling (Li *et al.*, 2009; McKenna *et al.*, 2010; Albers *et al.*, 2011; Koboldt *et al.*, 2012). However not all programmes perform equally well, and all have their own potential shortfalls (Shen *et al.*, 2010; Albers *et al.*, 2011; Nielsen *et al.*, 2011; Wang *et al.*, 2011). There is much variation between these different analysis pipelines, which I explore in chapter 3. Therefore, the selection of the appropriate pipeline will have a large effect on the potential of a study to identify causative variants.

Furthermore, variant callers often identify many thousands of variants (See chapter 4) per patient. Therefore one of the challenges in NGS studies is how to decide which of these variants are responsible for disease (Ng *et al.*, 2009; Ng *et al.*, 2010c; Zhi and Chen, 2012). This presents a major challenge, and as such many "variant filtering" procedures have been developed (Ng *et al.*, 2010b; Erlich *et al.*, 2011). Filtering methods include the selection of variants present only in cases (Bamshad *et al.*, 2011), removing variants identified outside of the target regions (Dickinson *et al.*, 2011), and incorporating a prediction programme to assess the potential impact of the variant on protein function (Bamshad *et al.*, 2011). However, many of these filtering methods imply a choice of genetic/biological model, which may not always be known. A detailed discussion of the approaches to variant filtering that I adopted is presented in the relevant chapters of this thesis.

NGS methods have been successfully applied to a range of monogenic, Mendelian diseases (Ng *et al.*, 2009; Ng *et al.*, 2010b; Bamshad *et al.*, 2011; Gilissen *et al.*, 2011). Due to the large capacity and cost requirements of whole genome sequencing (Teer and Mullikin, 2010; Gilissen *et al.*, 2011), whole exome sequencing is currently the more popular approach (Gilissen *et al.*, 2011; Smith *et al.*, 2011). The advantages and disadvantages of targeted capture approaches are discussed in chapters 3, 4, and 6.

1.7. Specific aims

The aim of my PhD was to use NGS methods to identify rare and potentially disease causing variants in various diseases, in particular CVMs. First, I implemented analysis pipelines, in the process becoming the first researcher in Newcastle University's Institute of Genetic Medicine to analyse NGS whole exome data. Realising the constant questions regarding the adequacy of sequencing quality, coverage, and analysis methodology that arise during NGS studies - particularly when nothing is found in a promising family and some estimate of the merit of investing further effort is required - I developed and tested a method to assess the performance of different NGS variant calling pipelines and the adequacy of a given set of sequencing output. I then applied the methods I had developed in two clinical contexts, attempting to identify rare variants that were disease-causing. The first used NGS methods to identify variants responsible for three Mendelian diseases (Atroventricular Septal Defects, Dilated Cardiomyopathy, and Hereditary Sclerosing Poikiloderma), and the second used NGS methods to identify potentially disease causing variants in a group of patients suffering from sporadic, non-Mendelian CVMs. In the second project, recognising the bioinformatics challenges that remain regarding the accurate calling of indels, I tested different analysis pipelines and evaluated their performance using the methods I developed earlier in my work.

Chapter 2. Methods

2.1 Methods overview

With regard to the bioinformatics analyses that were my principal focus, methods were specific to each sub-project and are described in the respective chapter. In this section, I provide an overview of those methods (chiefly laboratory methods) which were employed throughout my thesis to generate the data I was responsible for analysing. Experimental strategy for the laboratory work was decided by my supervisory team and as work progressed, was informed by the results I generated. Unless specified otherwise, the information in this chapter, which is chiefly provided for reference, was provided by Dr. Thahira Rahman and Mr. Rafiqul Hussain (University of Newcastle), who were the individuals responsible for carrying out the "wet lab" work. However, all data analysis and bioinformatics analysis was performed by myself.

2.2 Samples, target enrichment and sequencing

In chapter 4, I analysed data from cases in three pedigrees suffering from Mendelian diseases. In chapter 5 I analysed cases from the HeartRepair project, which consisted of 133 unrelated individuals suffering from various sporadic CVMs characterised by hypoplasia of one or other of the ventricles of the heart. The library preparation, target enrichment and sequencing varied between these two studies and will be discussed in more detail in this methods chapter.

2.2.1 Whole exome sequencing of Mendelian family samples (studied in chapter 4)

Blood and saliva samples were collected from members of three pedigrees. In each pedigree predisposition to a different disease appeared to segregate following an autosomal dominant mode of inheritance. Family trees and discussion of the phenotypes are provided in chapter 4.

Five micrograms of genomic DNA was extracted from pre-existing samples. A Covaris instrument (http://covarisinc.com/; University of Newcastle) was used to shear the genomic DNA into fragments ranging in size from 150 to 200bp. Fragmented samples were then assessed on a DNA1000 lab chip (Agilent) on an Agilent Bioanlyser to determine whether they were in the correct size range. The next step involved the end repair and adenylation of the fragments, after which Illumina adapters were added. The ligated products were then cleaned up to select only those DNA fragments which have adapters on both ends. Then in the following PCR stage, additional sequences were added to the ends of the adapters so that the final amplified templates contained sequences to enable hybridisation with primers bound to the flow cell surface for cluster generation (http://www.genomics.agilent.com).

Enrichment then proceeded using either the 38Mb or 50Mb SureSelect Human All Exon Capture Kit from Agilent Technologies (http://www.agilent.co.uk). Each enrichment kit consisted of custom biotinylated SureSelect oligonucleotides (also known as baits; http://www.genomics.agilent.com). The RNA bait-DNA hybrids were captured on streptavidin-coated magnetic beads. Amplification by PCR followed, after which the target samples were loaded onto the Illumina Genome Analyser IIx for paired end sequencing.

2.2.2 Targeted sequencing of unrelated cases of CVM (studied in Chapter 5)

Peripheral blood or saliva samples were collected from patients of European ancestry. Unrelated patients were recruited from four sources: (1) CONCOR (National Registry and DNA bank of congenital heart defects), The Netherlands, n=59 (2) National Registry for Congenital Heart Defects (NR-CHD), Berlin, Germany, n=13, (3) The Institute of Human Genetics, Newcastle University, United Kingdom, n=48, and (4) the University Hospital Zürich, Zürich, Switzerland, n=13. Informed consent was obtained from all participants according to institutional guidelines after approval of local ethics committees. Probands were evaluated by history taking, review of medical records, physical examination, 12-lead electrocardiography and transthoracic echocardiography. Family members, preferably the parents of the proband (trios), were included wherever possible but were only evaluated by history taking and review of medical records. Newcastle patients and their parents (when available) were recruited in the Freeman Hospital, Newcastle upon Tyne, UK after ethics approval by the Northern and Yorkshire Multi-centre Research Ethics committee.

Genomic DNA was extracted from blood and saliva samples by either the Phenol-chloroform method or the Oragene kit (DNA Genotek, Canada) respectively. For the NR-CHD patients genomic DNA was extracted by the Gentra Autopure LS automated DNA purifier (Gentra Systems, Minneapolis, USA) Genomic DNA samples were quantified by Qubit® Fluorometer (Invitrogen, Life Technologies) and their quality assessed on a Nanodrop (California, USA). 5 µg of gDNA was suspended in 1X TE buffer and made up to 100 µl. This was sonically sheared to fragments with average size of 200bp on the Covaris S2 system (Covaris, Massachusetts, USA). After fragmentation, samples were purified using QIAquick columns from a QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in Qiagen buffer EB. The size of fragments was assessed on the 2100 Bioanalyzer using DNA1000 chip.

The indexed paired-end libraries were prepared using reagents from the Illumina Genomic DNA Sample preparation kit and Multiplexing sample preparation oligonucleotide kit. Nucleotide overhangs produced as a result of the shearing process were converted to blunt ends using Klenow enzyme, T4 DNA polymerase, Klenow enzyme and T4 PNK. After incubation, samples were purified using QIAquick columns from a QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in Qiagen buffer EB.

Double stranded, blunt, phosphorylated DNA fragments were adenylated at their 3'ends in a reaction mix containing Klenow buffer, dATP and Klenow fragment (3' to 5' exo minus). After incubation, samples were purified using QIAquick MinElute columns from a MinElute PCR Purification Kit (Qiagen, Hilden, Germany) and eluted in Qiagen buffer EB.

Illumina multiplex paired-end adapters (Multiplexing Sample Preparation Oligonucleotide Kit, PE-400-1001, Illumina, San Diego, USA) were ligated to the adenylated DNA fragments in a reaction mix containing DNA sample, ligase buffer and Illumina multiplex paired-end adapter oligo mix and DNA ligase. After incubation, samples were purified using SPRI beads (Qiagen, Hilden, Germany) and eluted in Nuclease-free water. Quality and concentration of adapter ligated DNA fragments were checked on the 2100 Bioanalyzer using DNA1000 chip.

Four hundred and three (Appendix table 2.1) genes that were linked to cardiovascular malformations in humans or animal models were selected for capture and sequencing (further details on gene selection strategy are presented in Chapter 4). DNA sequences were downloaded from the UCSC Genome Browser and the coordinates of genomic sequences were based on NCBI genome build 36. The target regions encompassed 5152 exons and 1.68 Mbp. In these regions, 50406 unique capture probes were designed using the eArray algorithm following the manufacturer's instructions and with 5× tiling frequency. The biotinylated 120-mer cRNA probes were synthesised by Agilent Technologies (SureSelect target Enrichment System).

Target enrichment was performed using a SureSelect Custom Target Enrichment Kit (Agilent, California, USA). Libraries were made up to147 ng/µL ready for *in-solution* hybridisation with the custom SureSelect biotinylated cRNA oligonucleotides. To reduce non-specific hybridization, human Cot-1 DNA (2.5 µL) and 0.6 µL of a custom-made oligonucleotide block 3 pool containing equimolar concentrations (100 µM each) of four oligonucleotides.

The biotinylated cRNA-DNA hybrids were separated from the hybridisation mixture using Dynabead M-280 Streptavidin (Invitrogen, California, USA), washed and cRNA baits were digested following recommended protocol. Enriched fragments were purified using QIAquick MinElute PCR purification columns from the MinElute PCR Purification Kit (Qiagen, Hilden, Germany).

A post-hybridisation PCR was performed to barcode adapter ligated fragments and to selectively amplify enriched samples. The amplification reaction mix was prepared using Captured DNS fragments, Phusion High Fidelity DNA Polymerase master mix (Finnzymes, Finland), PCR primer InPE 1.0, PCR primer and PCR primer Index (Illumina Multiplex PCR kit) and incubated in a thermocycler. Quality and concentration of enriched DNA fragments were accessed on 2100 Bioanalyzer using the High-Sensitivity DNA kit.

Five samples were pooled at equimolar concentration and sequenced in a single lane of the Illumina GAIIx or Illumina HiSeq2000 sequencer with separate priming for reading the 6 nucleotide index sequences. For clustering on the GAIIx version 2 cluster kits were used on the GAIIx cluster station and for clustering on the cBot cluster station (for HiSeq 2000) version 3 HS cluster kits were used. Version 4 sequencing reagents were used for GAIIx and SBSv2 sequencing reagents were used for Hiseq2000. Base calling was performed with BclConverter-1.7.1 and subsequent use of Illumina's GAPipeline-1.5.1 (GAIIx) or CASAVA-1.7 pipeline (HiSeq2000). Demultiplexing was performed with a custom perl script allowing for one mismatch with the expected index sequence. For an initial assessment of coverage in targeted regions, sequences were aligned to the hg18 genome with bowtie v0.12.7, and only samples with at least 20x coverage in at least 80% of the targeted regions were used. In case of insufficient coverage, sequencing was repeated once or twice and resulting fastq files were merged. This was done at the AMC, Amsterdam.

2.3 Data analysis

2.3.1 Computers

Originally, the work was begun using a Dell PowerEdge 2970 system including 2 AMD Opteron HE 212 (2.0GHz) processors. This system had 12 cores, 32GB ram and 8TB storage.

The institute then acquired a cluster including a Dell PowerEdge R510 headnode, and 16 nodes housed in 4 Dell C6100 chassis. The headnode consisted of 2 Intel Xeon E5503 (2.0GHz) processors 8 cores and 12GB ram. The 16 additional nodes comprised 2 Intel Xeon E5640 (2.67GHz) processors, 8 cores, 48GB ram, and 150GB scratch space.

The cluster was later upgraded to a cluster including the original headnode, the original 16 nodes mentioned above, 4 new nodes, and a login node. The new login node comprised a Dell PowerEdge C1100 system with 2 Intel Xeon E5640 (2.4GHz) processors, 8 cores, and 24GB ram. The 4 new nodes are housed in a Dell C6100 chassis and each node consists of 2 Intel Xeon E5640 (2.66GHz) processors, 8 cores, 96GB ram and 1TB scratch space.

In total the cluster provided 20 nodes, 37TB lustre storage (via 2 Dell PowerEdge R510 processors), 20TB direct cluster NFS storage and 27TB attached NFS storage (via 2 Dell PowerEdge R715 processors). It runs a Scientific Linux release 6.3 operating system, and the OGS/GE 2011.11p1 batch-queuing system.

2.3.2 Scripting

In all chapters, I constructed analysis scripts using the Perl programming language (http://www.perl.org/), versions 5.10.1, 5.12.4, and 5.16.1. Perl scripts were used throughout my PhD to perform a range of tasks, and some of the more relevant tasks will be highlighted below.

I combined many of the alignment and variant calling programmes into single pipelines throughout my PhD. These created single analysis pipelines which, on specification of input fastq files and the desired criteria, would perform the alignment, variant calling, and the variant prediction stages of the pipeline. Appendix script 2.1 is an example of a script that I used to run MutationTaster.

In chapter 3 I developed a novel method to assess the performance of an entire next generation sequencing experiment. Due to the sheer volume of data contained in both the HapMap and 1000 Genomes databases, I developed scripts which could be used to extract only the on-target SNPs which were required for my method to assess the performance of the analysis pipelines (Method described in more detail in chapter 3).

More importantly though, a script had to be developed which given a list of reference SNPs and a VCF file, would calculate the sensitivity and specificity of the variant calls (See chapter 3 for more details on the method). Due to the length and complexity of the script it will not be made available here, but in the publication of the method.

Many aspects of chapter 4 involved the development and use of complex scripts for data analysis. In addition to the scripts which were used to run the analysis pipelines, this chapter also required the use of various scripts to perform all of the downstream data analysis steps. For example, given a list of identified variants I developed and used scripts that would remove all of the off-target variants and select those not found in the dbSNP database as well as removing those variants found in a control list.

Many of the scripts in chapter 5 performed similar functions to those in chapter 4. For instance, removing the off-target variants and those present within the control list (See methods in chapter 5). However, this chapter did require some additional scripts that were designed and implemented by myself. For example, I had to design a script to calculate the length of all the identified insertions and deletions (Appendix script 2.2). I also designed a script to identify the overlaps of indels identified by both the BWA-Dindel and BWA-GATK analysis pipelines (Appendix script 2.3).

2.3.3 Sequence analysis

Base calling was done using GERALD (cassava 1.6.0).

Alignment was performed using a series of programmes, specified in the relevant chapters. These included:

1) Bowtie v0.12.8 (Langmead *et al.*, 2009)
2) Bowtie2 v2.0.0beta6 (Langmead and Salzberg, 2012)
3) BWA v0.5.10 and v0.5.8 (Li and Durbin, 2009)
4) GMAP (GSNAP) v20120720 (Wu and Watanabe, 2005)
5) NovoAlign v2.07.13 (www.novocraft.com)
6) SSAHA2 v2.5.5 (Ning *et al.*, 2001)

Variant calls were also performed using a variety of programmes, which again are specified in the relevant chapters. These included:

1) Dindel v1.01(Albers *et al.*, 2011)
2) GATK v2.2.9 (McKenna *et al.*, 2010)
3) Samtools v0.1.8 and v0.1.18 (Li *et al.*, 2009)
4) Varscan v2.3.1 (Koboldt *et al.*, 2012)

All statistical analyses was performed using the R package (http://www.r-project.org/), versions 2.15.0 and 2.15.1.

2.3.4 Accessory programmes

Both a local version of Annovar (v506 and v510) and the online wAnnovar (http://wannovar.usc.edu/) version were used for variant annotations in chapter 4 and 5. Given a list of input variants, these two programmes annotate their functional affects and match them to public SNP databases using information form databases such as Ensembl (http://www.ensembl.org/index.html), 1000 genomes (www.1000genomes.org/) and the Exome Variant Server (EVS; http://evs.gs.washington.edu/EVS/) project.

MutationTaster (http://www.mutationtaster.org/) was used to predict the pathogenicity of variant calls.

Where applicable, BAM files were converted to fastq files using Bam2fastq v1.1.0 (http://www.hudsonalpha.org/gsl/information/software/bam2fastq).

The ngsqctoolkit v2.2.3 (Patel and Jain, 2012) was used to generate the graphs displaying average base coverage across the sequence reads in chapter 4.

The Picard v1.75 set of programmes was used in chapters 4 and 5 to remove duplicate reads from the sequence files.

In addition graphs were generated using the R package and SigmaPlot v11.

Chapter 3. Using population data for assessing next-generation sequencing performance

3.1 Aim

In the work outlined in this chapter, I describe the design of a simple, fast and effective method to calculate sensitivity and specificity of variant calls from NGS data. This is often done by comparing identified variants with those obtained using a reference technique such as a genotyping microarray. I designed a novel technique to calculate sensitivity and specificity using publically available population frequency data obtained from databases such as the HapMap and 1000 genomes databases.

I demonstrate that my method provides comparable results to those requiring microarray data. I compare different analysis pipelines used to identify single base substitutions and indels. As my method relies on allele frequencies obtained from public databases I also investigate the impact of using “incorrect” frequency data on sensitivity and specificity, and also explored the influence of sequence coverage.

3.2 Introduction

NGS technologies are often used for the identification of sequence variants predisposing to diseases that follow Mendelian inheritance patterns (Ng *et al.*, 2010a; Ng *et al.*, 2010b; Wang *et al.*, 2010a; Bamshad *et al.*, 2011; Gilissen *et al.*, 2011). Here I will define a variant as a deviation from a reference sequence. In particular, the sequencing of material enriched for exonic sequences has been successful in many cases but failed to identify the causative variants in others (Ng *et al.*, 2009; Bamshad *et al.*, 2011; Gilissen *et al.*, 2011). These successes and failures are described in more detail in chapter 4. In fact less than 50% of such studies are able to successfully identify the disease causing variant (Gilissen *et al.*, 2011). Such apparent failures may have many causes but also focus attention on the desirability of simple measures to assess the performance of different sequencing and analysis pipelines. The ideal analysis pipeline would have a high probability of identifying a variant, while maintaining a low number of falsely identified variants reducing the amount of work needed for validation.

Therefore, I developed a simple and flexible approach for assessing the performance of a whole exome or genome sequencing experiment. The method allows for the assessment of an entire sequencing study, from sample preparation through to variant calling. Here, the focus was on the detection of single base sequence variants as opposed to changes in copy number or large rearrangements. A common approach is to compare the identified variants with variants known to be present or absent, by using, for example, genotyping microarrays (Ng *et al.*, 2009). This allows for the probability of a specific variant at a given position in an individual to be considered as either 0 or 1. However, this obviously requires both sequence data and microarray data to be available which will increase the cost of an experiment, and is therefore not always a feasible approach.

An alternative approach, as is proposed here, is to use changes where the probability of occurrence in a specific sample can be ascertained. This allows the probability of a variant present at a specific position to assume values other than 0 or 1. My approach formalises this method by using sites known to be polymorphic in the human population. This can also be seen as an extension of methods that rely on quality criteria such as the number of variants found in sites known to be polymorphic (Marth *et al.*, 2011; Challis *et al.*, 2012).

The results of such comparisons can be summarised in many ways, such as assessing the number or proportion of variants that have previously been reported in databases such as HapMap (www.hapmap.org) or the 1000 genomes (www.1000genomes.org). Since the focus is on a dichotomous outcome I use here the probabilities of identifying the variant given that the variant is really present and of finding no variant at a site where none is present. I refer to these probabilities as sensitivity and specificity respectively.

I describe the method in detail below. In the results section I have compared my method to the method of calculating sensitivity and specificity using microarray data and explored possible applications of the proposed method in NGS studies.

3.3 Method

I designate with M the presence of a variant allele and with D the detection of a variant allele. Correspondingly $\overline{M}$ and $\overline{D}$ represent the absence of a variant allele and the not detecting a variant allele. For an autosomal locus we have:

$$p(D_i) = p(D_i \mid M_i M_i)p(M_i M_i) + p(D_i \mid M_i \overline{M}_i)p(M_i \overline{M}_i) + p(D_i \mid \overline{M}_i \overline{M}_i)p(\overline{M}_i \overline{M}_i)$$
$$p(\overline{D}_i) = p(\overline{D}_i \mid M_i M_i)p(M_i M_i) + p(\overline{D}_i \mid M_i \overline{M}_i)p(M_i \overline{M}_i) + p(\overline{D}_i \mid \overline{M}_i \overline{M}_i)p(\overline{M}_i \overline{M}_i)$$

Assuming Hardy-Weinberg equilibrium we obtain for the genotype frequencies $p(M_i M_i) = f_i^{\,2}$, $p(M_i \overline{M}_i) = 2 f_i (1 - f_i)$ and $p(\overline{M}_i \overline{M}_i) = (1 - f_i)^2$, where f_i designates the frequency of the variant allele. We further designate with *s* the sensitivity $s = p(D \mid M)$ and with *u* the specificity, $u = p(\overline{D} \mid \overline{M})$ and obtain for the remaining terms: $p(D_i \mid M_i M_i) = s(2-s)$, $p(D_i \mid M_i \overline{M}_i) = s + (1-s)(1-u)$, $p(D_i \mid \overline{M}_i \overline{M}_i) = 1 - u^2$, $p(\overline{D}_i \mid M_i M_i) = (1-s)^2$, $p(\overline{D}_i \mid M_i \overline{M}_i) = (1-s)u$ and $p(\overline{D}_i \mid \overline{M}_i \overline{M}_i) = u^2$.

We treat all sites as independent and assume that detection probability for one site is independent from that for another, thus for an individual the likelihood is

$$l(s,u) = \prod_{i \in S_D} p(D_i) \prod_{j \in S_{\overline{D}}} p(\overline{D}_j)$$

Where the index S_D represents the set of sites where a variant was detected and $S_{\overline{D}}$ the sites were only the reference was observed.

When several individuals are analysed and we assume that their genotypes are independent the likelihood of the whole group of individuals can be described as

$$l = \prod_{k=1}^{K} l_k(s,u)$$

3.4 Materials

3.4.1 Sequence and genotype data

Targeted whole exome sequencing was carried out for 31 (12+19) samples using an Illumina Genome Analyser IIx. Agilent 38Mb target positions were obtained from Agilent, and the human genome (Build36.1) was used as reference sequence (http://genome.ucsc.edu/).

Genotype chip data were available for 19 of the 31 samples, comprising a total of 557124 SNPs on the Illumina 660 genotype chip (https://my.illumina.com/)**.** 10762 of these SNPs were located within the target regions.

3.4.2 Comparison of array and sequencing data

As described in the introduction, my analysis focuses on the ability of detecting variants, therefore I assessed at any position whether a variant was detected or not. The sensitivity is defined as the number of sites in which both sequencing and microarrays detected a deviation from the reference sequence divided by the number of sites where a variant was detected by using the microarrays. Correspondingly the number of sites where both methods detected no deviation from the reference sequence divided by the number of sites where the microarray detected only the reference residue was used as an estimate of the specificity.

3.4.3 Selection of polymorphisms

The HapMap database was used to obtain allele frequencies in the calculation of sensitivity and specificity for the single base substitution variant calls. The HapMap database consists of 4083713 SNPs in total (CEU population, build 36, downloaded 28 October 2010). 10165 overlap with the on target (Agilent 38Mb whole exome targets) genotype chip SNPs. Allele frequency data was obtained for each of these SNPs from the HapMap database.

For the calculation of sensitivity and specificity of indel calls, allele frequencies were obtained from the 1000 genomes database. As indel positions were not typed on the microarray, I used all the indel positions from the 1000 genomes database which were located in the target regions (Agilent 38Mb whole exome targets). This provided 8365 on target indels from the 1000 genomes database for which frequency information was also available.

3.4.4 Sequence analysis

For the identification of single base substitutions, reads were aligned to the reference using the following aligners: Bowtie (Langmead *et al.*, 2009), BWA (Li and Durbin, 2009), GSNAP (Wu and Nacu, 2010), NovoAlign (http://www.novocraft.com/), SOAP2 (Li *et al.*, 2008b), SSAHA2 (Ning *et al.*, 2001). Variants were identified using either Varscan (Koboldt *et al.*, 2009) or Samtools (Li *et al.*, 2009). Unless specified otherwise, the default parameters were used for each program. Coverage was assessed from the Pileup files. Coverage depth was varied by sampling with replacement from the SAM files.

Twelve of the samples were used for indel analysis. Reads were aligned to the human genome reference (Build 37, hg19) sequence using Bowtie2 (Langmead and Salzberg, 2012), BWA and NovoAlign. Indels were identified using Samtools, Dindel (Albers *et al.*, 2011) and GATK (McKenna *et al.*, 2010; DePristo *et al.*, 2011). Due to the large number of windows produced by Dindel, a minimum threshold of 7 reads covering each indel variant was applied. For all other indel callers, the default parameters were used.

3.5 Results

In this section I first explore some applications of the proposed method, then compare its results with estimates generated using the genotypes obtained through the microarray as the true genotypes. The method I propose here assumes that the allele frequencies in the population from which the samples were drawn are known. At the end of this section I explore the effects of using incorrect allele frequencies.

3.5.1 Pipeline comparisons

Figure 3.1 compares the sensitivity and specificity estimates achieved using the different alignment and variant calling programmes for the detection of single nucleotide substitutions. The values are based on 31 samples and represent the median and the upper and lower quartiles. For 8 samples SSAHA2 failed to produce results and generated the messages “error: memory allocation failed cannot allocate memory” or “error: memory allocation for array of fasta structures failed cannot allocate memory”.

Interestingly, the alignment programmes appear to have a stronger effect on sensitivity, while the variant calling programmes appeared to effect specificity more strongly. All aligners yield similar specificities when used in combination with Samtools and the NovoAlign-Samtools and BWA-Samtools pipelines provided the highest sensitivity values. Therefore I used the NovoAlign aligner and the Samtools variant caller as the standard pipeline for all of the subsequent analyses. The estimates for the specificity using Samtools with any aligner were over 0.998. Conversely, the specificity fell below 0.996 when using Varscan.

Figure 3.2 shows the results for different indel calling pipelines. The number of positions considered makes estimation of specificity problematic. Sensitivity was poor compared to the value for single base substitutions across all pipelines and it was shown that Samtools had the lowest sensitivity, while GATK performed best. The best performing pipeline was BWA-GATK. The average estimates for sensitivity using the Bowtie2-GATK and Novoalign-GATK pipelines are 0.354 and 0.337 respectively.

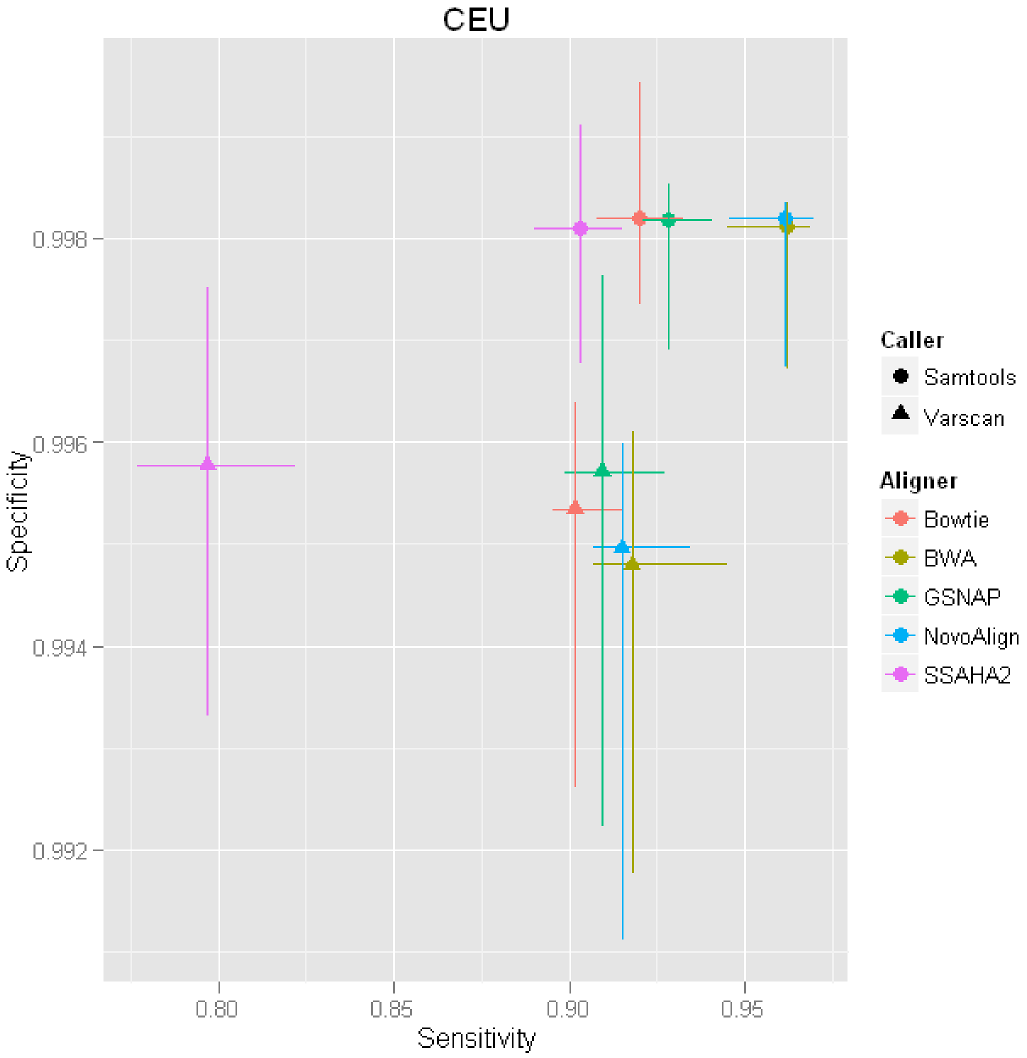


Figure 3.1. Specificity and sensitivity of different analysis pipelines used to call single base substitutions.

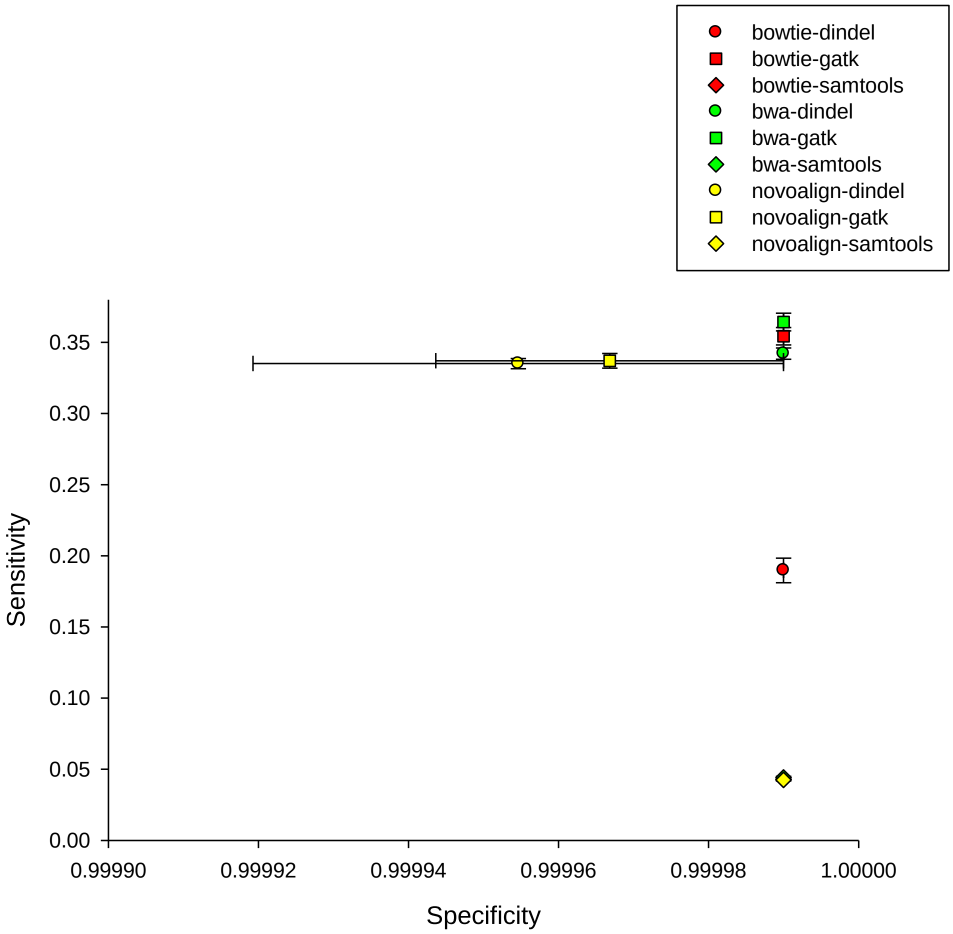


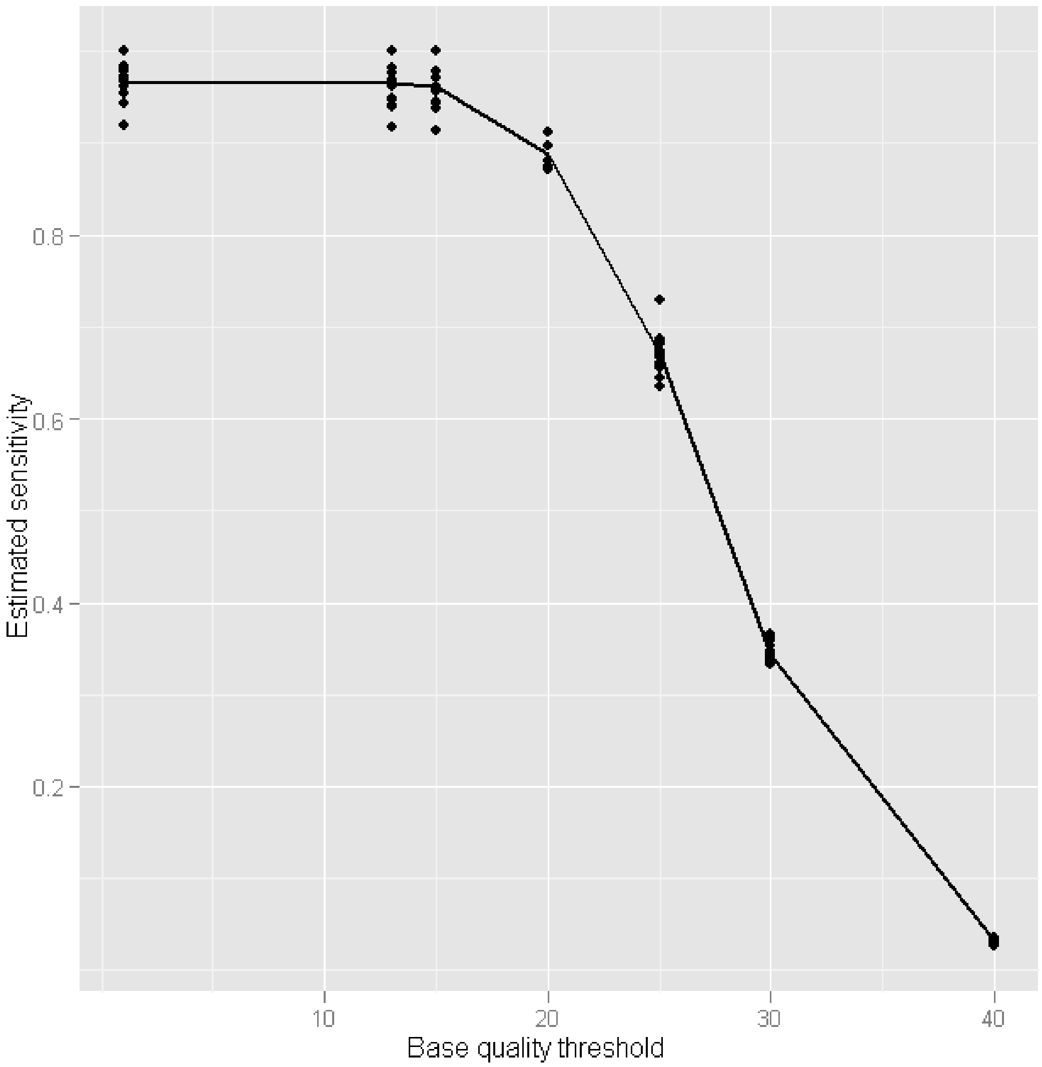
Figure 3.2 Specificity and sensitivity for indel calling pipelines.

3.5.2 Parameter selection

Each alignment and variant calling programme has a range of parameters which can be set by the user. In general default values are provided but these may not always be appropriate and altering these parameters can have a marked effect on sensitivity and specificity. Therefore, I used my method to explore the effects of choosing different variant calling parameters. To test this I used the NovoAlign-Samtools pipeline but used a range of base quality thresholds for variant calling (Figure 3.3A and 3.3B). This is the minimum base quality at the position, for a read, required for that read to be included in the variant call for that position.

Figure 3A shows that altering the base quality threshold used in variant calling has a dramatic effect on sensitivity, with a rapid drop when the values are set above 20. The effect on specificity (Figure 3B) is more modest and increasing the base quality threshold beyond 30 has only a limited effect.

3.3A


Estimated sensitivity
0.8
0.6
0.4
0.2
10
20
30
40
Base quality threshold

3.3B

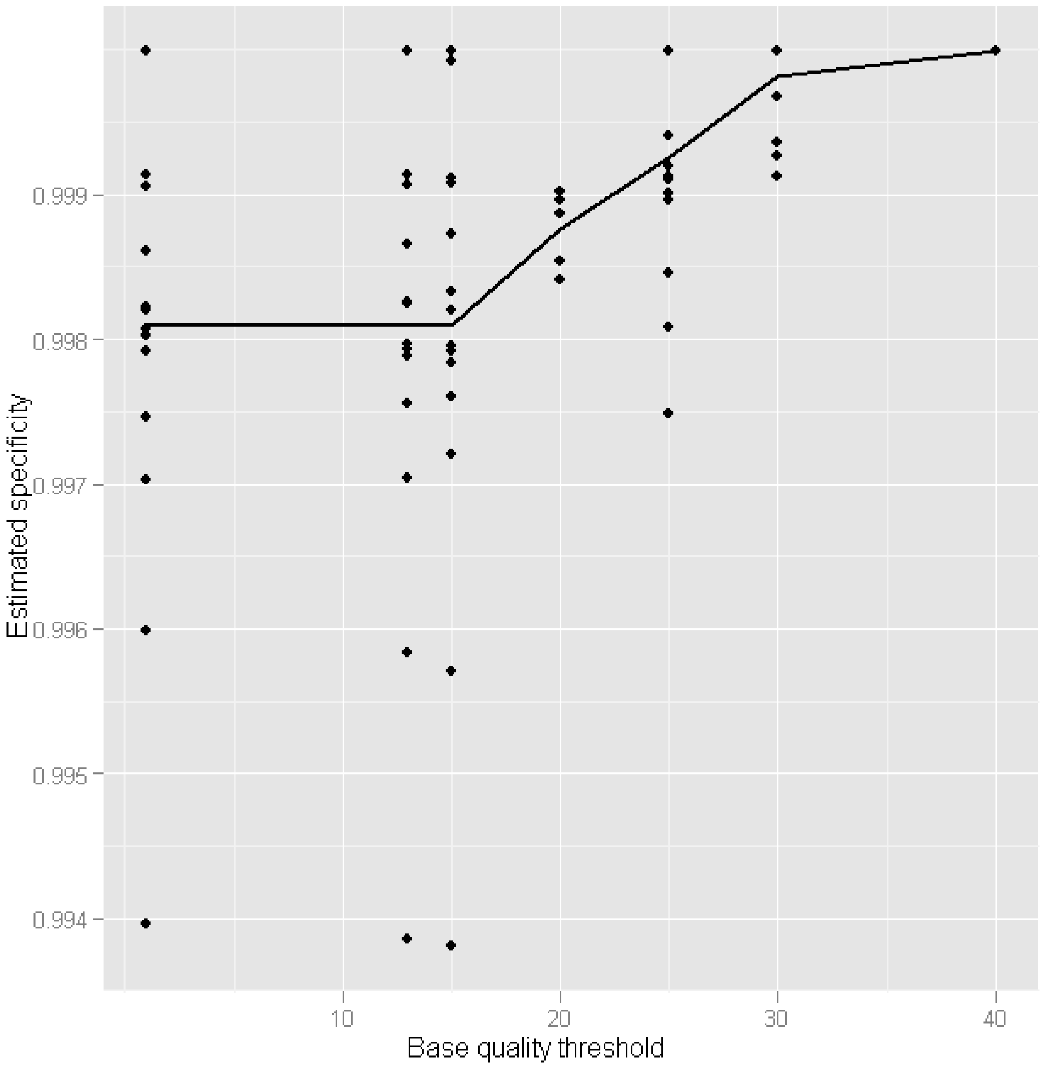


Figure 3.3 (A, B). The effect of parameter choice. Each point on the graph represents the value obtained for one sample using a particular base quality threshold. Figure 3.3A represents the influence of parameter choice on sensitivity, and Figure 3.3B represents the influence of parameter choice on specificity.

3.5.3 Coverage

Figure 3.4 explores the effects of average coverage on sensitivity. As expected, sensitivity increases as the coverage increases. This reflects the fact that at low coverage finding evidence for a variant generally becomes more difficult, thus leading to a low sensitivity. With the parameters used such a loss began to be evident when the average coverage was below 40 fold.

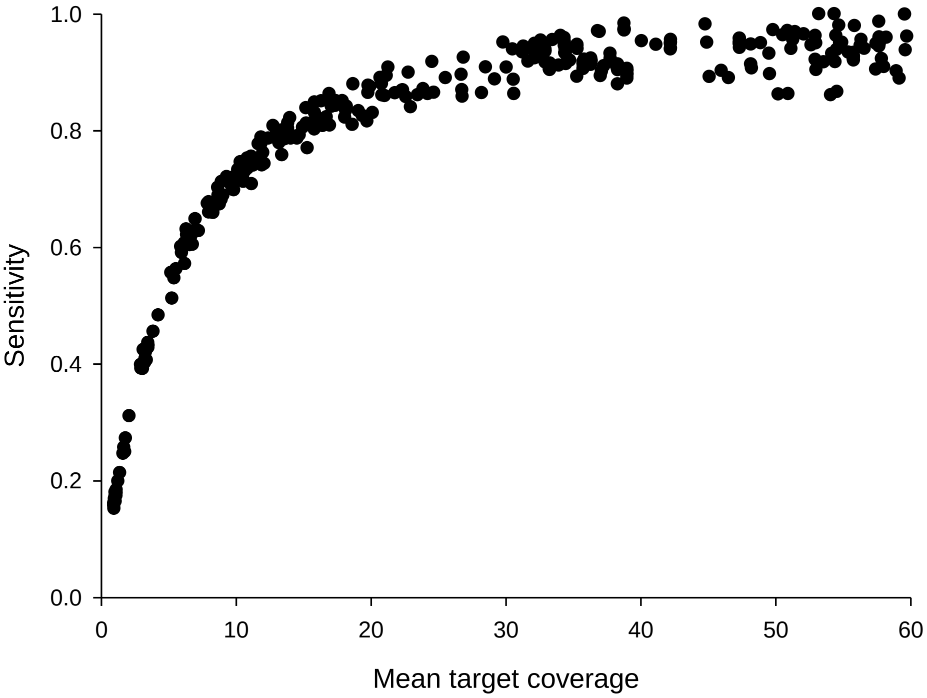


Figure 3.4. Influence of average coverage on sensitivity. Each point on the graph represents the sensitivity value calculated for a particular sample at a particular target coverage.

3.5.4 Microarray comparison

As stated above, both sequence data and microarray data were available for 19 of the samples. Table 3.1 compares the parameter estimates using the genotyping microarray data (Fourth column) and two different sets of allele frequencies (Second and third column). The values in the third column are derived from the genotyping results. Compared to microarrays both specificity and sensitivity estimates are slightly lower by the frequency method using the CEU population frequencies (Second column).

One possible reason for this is that the estimates are distorted because the HapMap CEU allele frequencies do not match the allele frequencies in my sample. Indeed the difference is smaller when the allele frequencies used are derived from the genotyping results of the 19 samples (Third column), and the sensitivity is even slightly higher than when calculated by comparing to the microarray data (Fourth column). This suggests that the sensitivity estimates are influenced by the choice of frequency data and should therefore be considered with care.

It should also be mentioned that specificity estimates are in the order of 0.999. These estimates however are based only on a limited number of polymorphic sites (10165 sites), suggesting that the ability to adequately assess changes in specificity will be limited. This is reflected in the correlation between the estimates obtained from microarrays and from population frequencies. While there is a good correlation between the estimates for the sensitivity (see Figure 3.5, R^2=0.71, P=4x10^{-50}), the correlation for specificity is rather poor although still significant (R^2=0.39,P=7x10^{-21}).

	Estimated from		
	CEU frequencies[a]	**Sample frequencies**[b]	**Microarray**[c]
Sensitivity **(95% CI**[d]**)**	0.962 (0.945-0.970)	0.979 (0.962-0.986)	0.984 (0.982-0.986)
Specificity **(95% CI)**	0.998 (0.997-0.998)	0.999 (0.999-0.999)	0.999 (0.999-1.00)

[a] Allele frequencies for the Hapmap CEU population

[b] Allele frequencies determined from the all the samples using the microarray genotyping results.

[c] Genotypes determined using the Illumina 660W chip

[d]: 95% confidence interval for the mean, determined by resampling.

Table 3.1. Mean sensitivity and specificity estimates. Represented are the estimates for the specificity and sensitivity of the NovoAlign-Samtools pipeline.

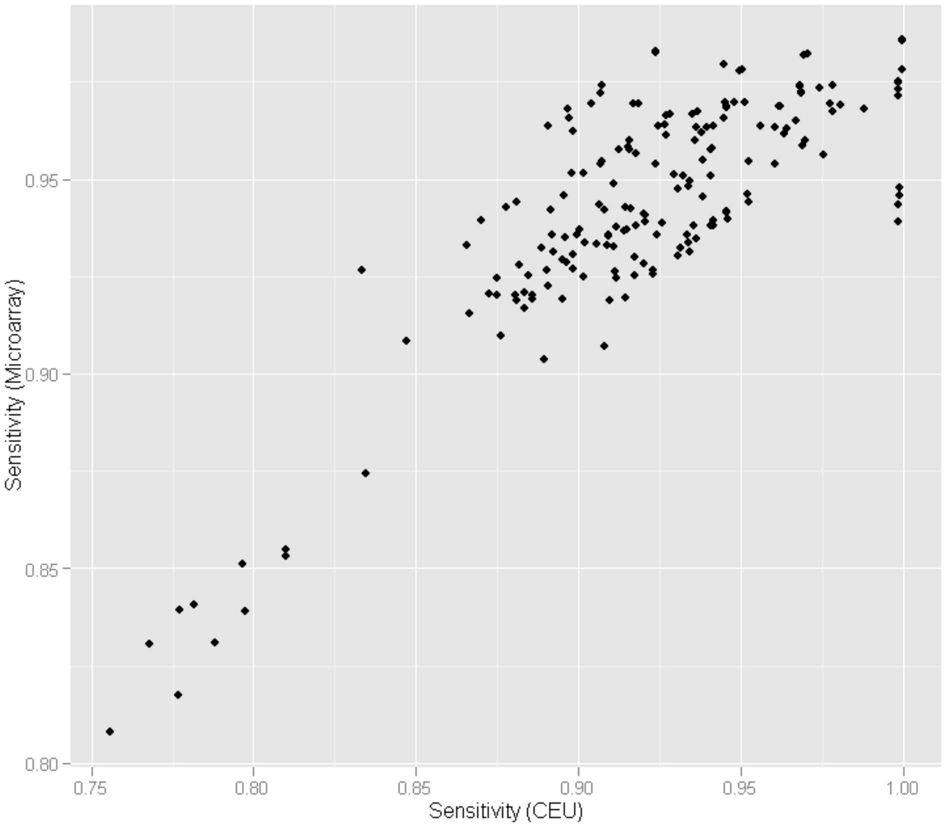


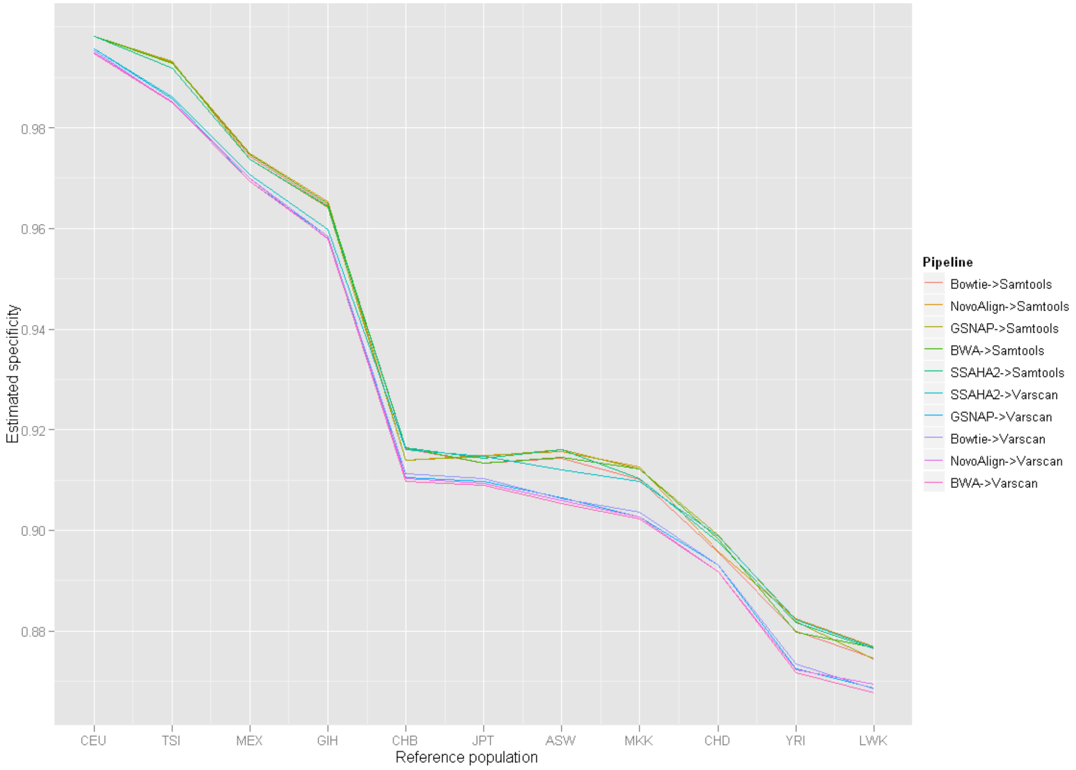
Figure 3.5. Correlation between sensitivity estimates from microarray data and using CEU population frequencies. The points represent the values for different individuals and analysis pipelines.

3.5.5 Influence of using different allele frequencies

The method proposed here uses allele frequencies in the calculation of sensitivity and specificity. However, allele frequencies vary between populations and the ethnicity of the individuals who provided a sample may not always be known. Therefore, I explored the influence of using different frequencies on sensitivity and specificity by using allele frequencies from all 11 HapMap populations (Figure 3.6A and 3.6B).

Figure 3.6 suggests that misspecification of population frequencies tends to lead to lower estimates of both sensitivity and specificity. However the lines connecting the values for the different pipelines calculated using different allele frequencies tend to be parallel. This indicates that the results are correlated ($p<0.01$ for all comparisons) suggesting that although the absolute values may vary, the order of the different pipelines will remain the same.

3.6A


Estimated specificity
0.98
0.96
0.94
0.92
0.90
0.88
CEU
TSI
MEX
GIH
CHB
JPT
ASW
MKK
CHD
YRI
LWK
Reference population
Pipeline
Bowtie->Samtools
NovoAlign->Samtools
GSNAP->Samtools
BWA->Samtools
SSAHA2->Samtools
SSAHA2->Varscan
GSNAP->Varscan
Bowtie->Varscan
NovoAlign->Varscan
BWA->Varscan

3.6B

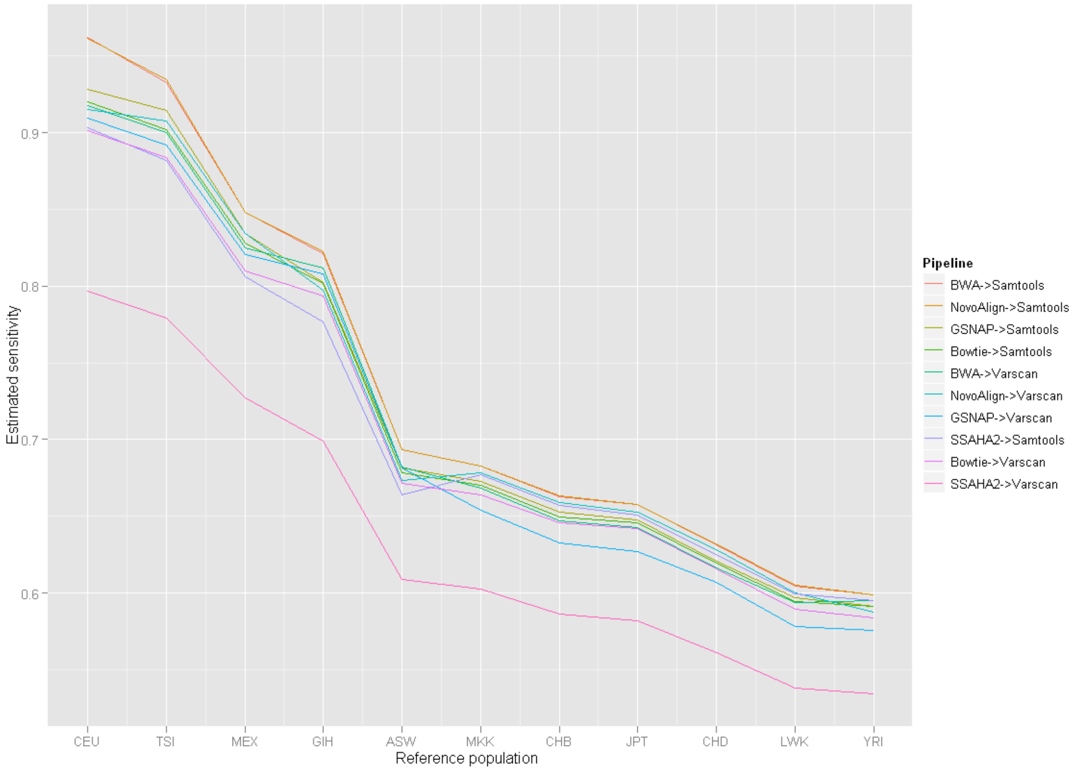

Estimated sensitivity
0.9
0.8
0.7
0.6
CEU
TSI
MEX
GIH
ASW
MKK
CHB
JPT
CHD
LWK
YRI
Reference population
Pipeline
BWA->Samtools
NovoAlign->Samtools
GSNAP->Samtools
Bowtie->Samtools
BWA->Varscan
NovoAlign->Varscan
GSNAP->Varscan
SSAHA2->Samtools
Bowtie->Varscan
SSAHA2->Varscan

Figure 3.6 (A, B). Effect of reference population misspecification on specificity (3.6A) and sensitivity (3.6B). The x-axis represents the different HapMAp populations for which allele frequency data were available. CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; TSI: Tuscan in Italy; MEX: Mexican ancestry in Los Angeles, California; GIH: Gujarati Indians in Houston, Texas; ASW: African ancestry in Southwest USA; MKK: Maasai in Kinyawa, Kenya; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; CHD: Chinese in Metropolitan Denver, Colorado; LWK: Luhya in Webuye, Kenya; YRI: Yoruban in Ibadan, Nigeria.

3.6 Discussion:

The approach presented here estimates two parameters, sensitivity and specificity, from NGS variant calls. I illustrated some of its potential applications by comparing analysis pipelines, variant calling parameters and exploring the effects of differences in coverage. Since both sensitivity and specificity are influenced by various steps including sample preparation, the sequencing itself and the bioinformatic pipelines, the procedure could be used to assess the performance of a sequencing experiment globally and could complement other commonly used approaches such as the assessment of base call quality or of coverage metrics. The main advantage of the method presented here is that it does not require a reference technique, such as genotyping using microarrays. However, it does rely on the availability of appropriate allele frequencies.

The use of allele frequencies has two consequences. The first is that compared to the situation where the presence or absence of variants is known, using a probability introduces a degree uncertainty that is reflected in a larger scatter of the estimates (see Table 3.1). The second is that it forces one to decide which set of allele frequencies to use. The choice of the allele frequencies from a specific population or, if available, from a particular subpopulation, disregards the possibility that individuals may represent a mixture from different populations. This problem could be avoided using a more complicated approach that considers, for example, the probability of belonging to a certain population, or of carrying certain haplotypes and perhaps allowing these probabilities to differ for different regions of the genome. I demonstrated that the misspecification of the population frequencies will influence the specificity and sensitivity values. However, figure 3.6 suggests that if two procedures have substantially different specificity or sensitivity values, the use of different allele frequencies will still tend to preserve the order of the different analysis pipelines.

One very important issue is the location of the polymorphisms used. Here I chose those included in the regions targeted by the enrichment procedure. However, since the practical interest here is to detect variants likely to cause disease, it would perhaps be a better choice to use all the polymorphisms in coding or non-coding regions. Another issue surrounds the type of polymorphism included in the analysis. Here I chose the polymorphisms represented in the microarray. These polymorphisms represent a selection based on criteria that probably includes the likelihood of being efficiently typed using microarray technology. This will probably result in avoiding certain types of polymorphisms, such as indels, and polymorphisms in certain locations such as gene regions with extreme base compositions. It is possible that sequencing experiments are accurate, or inaccurate, in exactly the same regions and this would lead to a bias in the sensitivity and specificity values. However, the proposed method allows for a comparison of different types of polymorphisms and I showed its application to the identification of indels. As expected both specificity and sensitivity appear to be lower than the values for single nucleotide substitutions. Although allele frequencies for polymorphisms not included in microarrays may not be accurate, my results are consistent with published studies that show that indel detection is still a challenging issue (Albers *et al.*, 2011; Bansal and Libiger, 2011). Albers *et al.* (2011) compared the false discovery rate of indel calls using Dindel, Varscan and SAMtools. Dindel achieved the lowest false discovery rate of 1.56%, while Varscan achieved the highest rate of false discoveries, 16.67%.

Since the interest here was on the detection of rare variants, I was able to dichotomise the outcome by scoring at each position whether a variant was present or absent. This leads effortlessly into the determination of specificity and sensitivity. However more complicated scenarios are possible, such as assessing the calling of each of the three possible genotypes defined by a variant/reference allele combination at each position. Here, however, I chose the more simple approach.

Studies frequently focus on sensitivity as oppose to specificity (Pattnaik *et al.*, 2012), however specificity is also very important as it may help to assess the amount of validation work that is required, which is closely related to false positive rate. A specificity of 0.99 and a frequency of deviations from the reference sequence in the order of 1 per 1000 sites would be expected to lead, on average, to approximately ten false to one true positive. Therefore, practically useful methods of variant detection should have specificities that are much higher than 0.99. Estimating such a parameter accurately will require examining a large number of sites. For example, simply counting false and true negatives, relying for example on microarray data would require at least 10 650 variant positions to establish the difference between a method that has specificity of 0.999 compared to one with a specificity 0.9999 with 80% power. In the present study I relied on sites that are known to be polymorphic, however I could also include sites for which there are no reported variants and assume a low minor allele frequency for all of these sites. This would increase the number of sites used and improve the ability to estimate the specificity.

Since this procedure is quite simple it would be possible to use it to optimise analysis parameters, by integrating it into, for example, a variant caller so that it maximises the sensitivity while not allowing specificity to drop below a certain threshold. Such a procedure would benefit from the fact that the order appears to be insensitive to the choice of population. This would allow for an estimation of the amount of validation work required and the likelihood that a change of interest can really be identified, and can guide the design of future experiments.

In summary, I have developed a method to assess the performance of an entire exome NGS experiment. The major advantage of my method is that it does not require the use of a reference technique, but calculates sensitivity and specificity values using freely available frequency data from databases such as HapMap or the 1000 genomes. The proposed method is simple, and fast to implement but still produces sensitivity and specificity values comparable to those calculated using microarray data. Therefore, such a method could be simply used to inform the choice of analysis pipeline, analysis parameters or even of experimental protocol.

Chapter 4. Exome sequencing to identify the causative variants in three diseases showing transmission consistent with Mendelian inheritance

4.1 Aim

In the work outlined in this chapter, next-generation sequencing was used to identify potentially causative variants in three pedigrees where disease segregates in a Mendelian fashion consistent with a monogenic cause. Whole exome sequencing was carried out using an Illumina Genome Analyser IIx for both affected and unaffected individuals. The NovoAlign aligner and Samtools variant caller were originally used to identify both single base substitutions and indels, however on completion of chapter 3 I identified the BWA-Dindel pipeline as being a better pipeline for identifying indels. Therefore this was used as an additional pipeline for indel identification in the three families.

Various filtering steps were used to identify the most likely variant resulting in disease from amongst the many thousands of variants which were identified. I identified potentially disease causing variants in plausible candidate genes for disease in the pedigrees where cases presented with Dilated Cardiomyopathy and Hereditary Sclerosing Poikiloderma. There are various reasons why I was not able to identify any potentially disease causing variants in plausible candidate genes in the pedigree where cases presented with Atrioventricular Septal Defects, which will be discussed in this chapter.

4.2 Introduction

NGS methods have proved very successful in the search for the genetic causes of disease, particularly where disease follows a Mendelian inheritance pattern and appears to be monogenic (Ng *et al.*, 2009; Ng *et al.*, 2010b; Bamshad *et al.*, 2011; Gilissen *et al.*, 2011). Although sequencing the entire genome is a feasible option (Gilissen *et al.*, 2011), its uses are limited by the large capacity and cost requirements (Teer and Mullikin, 2010; Gilissen *et al.*, 2011). Instead, exome sequencing is a more popular approach (Gilissen *et al.*, 2011), because it is cheaper but still allows for the identification of variants within the coding portions of genes (Ng *et al.*, 2009; Bamshad *et al.*, 2011).

Although exome sequencing does not assess the potential impact of non-coding variants (Bamshad *et al.*, 2011), less than 1% of the identified variants in Mendelian disease have been found in non-coding regions (Ng *et al.*, 2008). Furthermore, many of the non-synonymous variants which have been identified in coding regions have been shown to be deleterious (Kryukov *et al.*, 2007; Stenson *et al.*, 2009; Ng *et al.*, 2010b; Bamshad *et al.*, 2011; Kiezun *et al.*, 2012), and the large effect of causal mutations in Mendelian diseases suggests that the variants are largely coding (Ng *et al.*, 2010c).

On the other hand, it is possible for the genetic variants causing some of these diseases to be present in the non-coding regions of the genome (Hirschhorn and Daly, 2005). For example, deep intronic substitutions in the *CDKN2A* have been shown to cause some types of melanoma (Harland *et al.*, 2001), and in other cases, intronic substitution events in the *SLC12A3* gene have been proposed to cause Gitelman's Syndrome (Nozu *et al.*, 2009).

Despite its popularity, the majority of exome sequencing projects have proved unsuccessful in identifying the causative variants responsible for disease (Gilissen *et al.*, 2011; Zhi and Chen, 2012). Gillisen *et al.* (2011) suggest that only 50% of studies involving rare, well defined Mendelian conditions are able to identify the genetic causes. One of the major challenges of exome sequencing is how to distinguish the disease causing variants from the non damaging, rare or even unique variants also present within an individual (Ng *et al.*, 2009; Ng *et al.*, 2010c; Erlich *et al.*, 2011; Stitziel *et al.*, 2011; Zhi and Chen, 2012). To this end various filtering methods have been developed to reduce the number of identified variants to encompass only those most likely to cause disease. The advantages and disadvantages of these methods will be discussed in more detail in the discussion section of this chapter.

In this chapter I aimed to employ exome sequencing to identify the causative variants for three different diseases segregating in a Mendelian fashion consistent with a monogenic cause. Sequence data was available for individuals from three pedigrees (Figures 4.2, 4.4, and 4.5). The affected individuals suffer from Dilated Cardiomyopathy (DCM), Atrioventricular Septal Defect (AVSD), and Hereditary Sclerosing Poikiloderma (HSP). Single base substitutions and indels were initially identified using the NovoAlign aligner and Samtools variant caller. However, the results obtained in chapter 3 indicated that the BWA aligner and Dindel variant caller would be more appropriate methods for indel identification. Therefore, the BWA-Dindel pipeline was employed as an additional indel calling pipeline. I include indel calls from both pipelines.

The identified variants were then filtered using methods commonly employed in the literature. After filtering I hoped to be able to identify potentially disease causing variants in plausible candidate genes of interest. Here I defined plausible candidate genes of interest as genes which appear interesting at first, based on current understanding of gene function and related diseases from databases such as OMIM. For example, potentially disease causing variants in genes that result in phenotypes similar to the ones in the present study. I will use this term throughout this chapter. For clarity the three different diseases will be introduced under their own subheadings below.

4.2.1 Dilated cardiomyopathy

Cardiomyopathies comprise a range of cardiac disorders affecting the heart muscle (Schonberger and Seidman, 2001; Towbin *et al.*, 2006). They are categorised based on their anatomic and haemodynamic attributes (Schonberger and Seidman, 2001). The two major forms of cardiomyopathy are dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (Parvari and Levitas, 2012). The affected individuals in this study suffer from DCM.

DCM is the most common form of cardiomyopathy, making up more than 80% of all cases (Schonberger and Seidman, 2001; Luk *et al.*, 2009). It results in the myocardial walls stretching and thinning which negatively affects ventricular function (Luk *et al.*, 2009). This diminished contractile function is one of the more serious haemodynamic features of the disease, and can lead to various complex, compensatory neurohumoral responses which later result in heart failure (Schonberger and Seidman, 2001). The clinical symptom of DCM is eventual heart failure often associated with arrhythmia and sudden death (Parvari and Levitas, 2012).

The severity of symptoms and survival varies considerably between patients and diagnosis is often made using non-invasive cardiac imaging (Schonberger and Seidman, 2001). Although the disease can manifest in early childhood, it is usually only identified later in life at which point it has often progressed to end-stage myocardial fibrosis (Schonberger and Seidman, 2001; Luk *et al.*, 2009).

Estimating the incidence and prevalence of individual conditions is difficult as the effect of disease may remain subclinical (Raju *et al.*, 2011). A ten year study conducted in the USA estimated the incidence of DCM to be ~6 per 100000 people per year, with a prevalence of ~36.5 per 100 000 people (Codd *et al.*, 1989). Whereas the Paediatric Cardiomypathy Registry estimates the annual incidence of DCM to be 1.13 cases per 100 000 people aged 18 years or younger (Parvari and Levitas, 2012).

The underlying causes of DCM are varied, and include both environmental and genetic factors (Tsubata *et al.*, 2000; Miyamoto *et al.*, 2001; Schonberger and Seidman, 2001; Luk *et al.*, 2009; Hazebroek *et al.*, 2012). Some of the principal causes include viral myocarditis (usually occult and unable to be proven at the time of presentation with DCM), thyroid disease (reversible with therapy), immunological processes, toxins (such as alcohol and heavy metals), drugs (notably anticancer chemotherapy) and infiltrative processes (Schonberger and Seidman, 2001; Hazebroek *et al.*, 2012). Nevertheless, the causes of up to 65% of cases remains unknown, a condition termed "idiopathic" DCM (Parvari and Levitas, 2012).

Among these idiopathic cases, genetic causation is prominent. Familial forms of DCM, which are mostly monogenic, account for 25-35% of idiopathic DCM cases (Miyamoto *et al.*, 2001; Towbin *et al.*, 2006; Luk *et al.*, 2009; McDermott *et al.*, 2012). Although autosomal dominant transmission accounts for about 70% of inherited DCM cases (Tesson *et al.*, 2000; Mahon *et al.*, 2005; Hazebroek *et al.*, 2012), three other modes of inheritance have also been identified, namely autosomal recessive, X-linked and mitochondrial inheritance (Luk *et al.*, 2009; Hazebroek *et al.*, 2012).

The discovery of the genes responsible for DCM has proven difficult, largely due to the presence of both substantial aetiological and genetic heterogeneity (Mahon *et al.*, 2005). Multiple genetic regions and genes have been shown to be involved (Tsubata *et al.*, 2000). Various methods have been used to identify these causative genes. These include using direct candidate gene sequencing in affected individuals (van der Zwaag *et al.*, 2012), association (Zarrouk Mahjoub *et al.*, 2012) and linkage studies (Yoskovitz *et al.*, 2012), or through a combination of approaches such as sequencing and association methods (Herman *et al.*, 2012).

Dellefave & McNally (2010) provide an electron micrograph of a cardiomyocyte which describes specific intracellular regions and the genes in these regions which have been shown to cause DCM (Figure 4.1). In addition, Hershberger *et al.* (2010) and Parvari & Levitas (2012) both provide detailed lists of known genes responsible for DCM. Table 4.1 is taken from Parvari & Levitas (2012), and describes all of the genes known to cause cardiomyopathies, in particular DCM. Most causative genes seem to encode for cytoskeletal and sarcomeric proteins, thus affecting the structure of the muscle (Dellefave and McNally, 2010; Hershberger *et al.*, 2010; Hazebroek *et al.*, 2012; Parvari and Levitas, 2012).

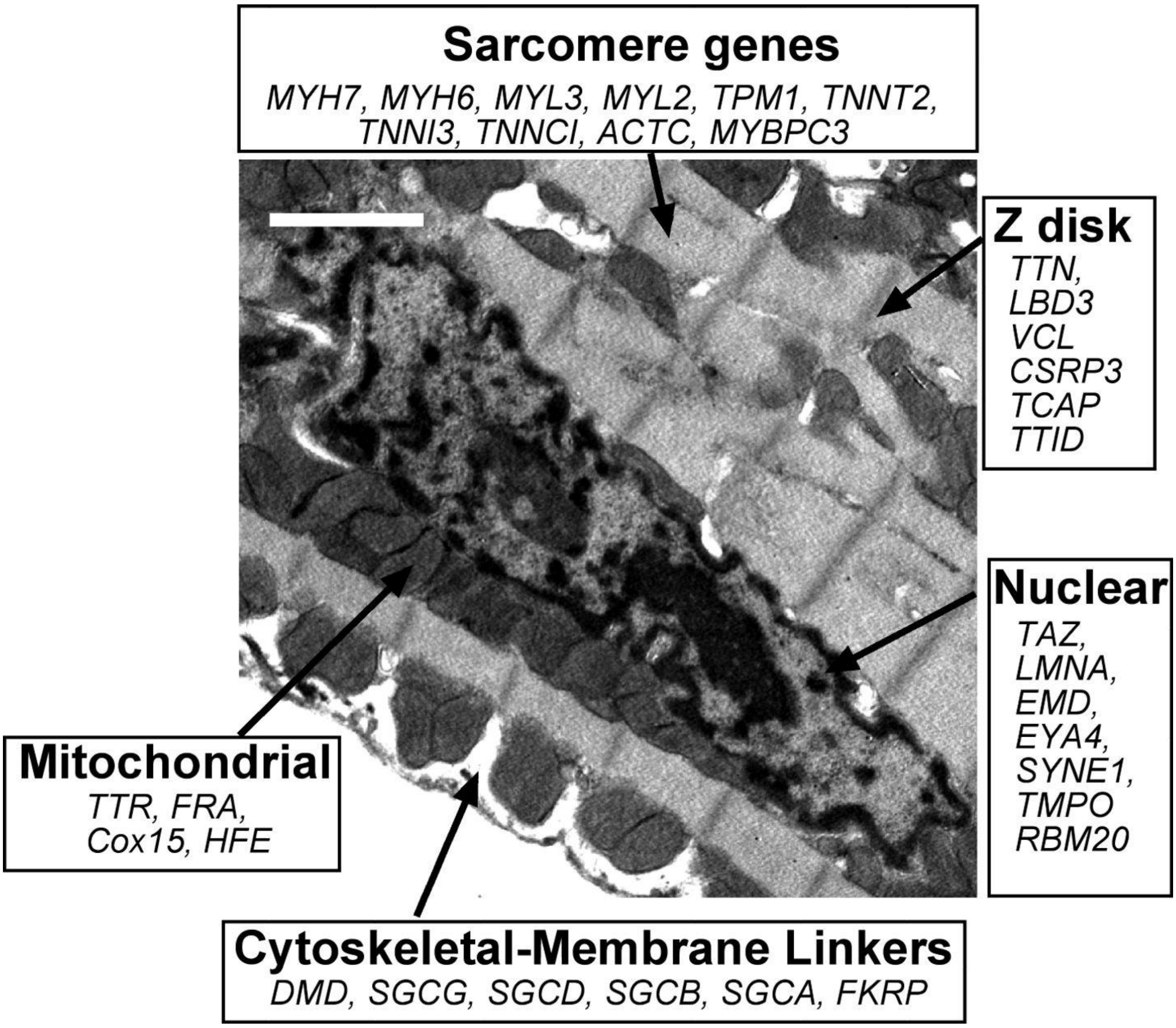

Sarcomere genes
MYH7, MYH6, MYL3, MYL2, TPM1, TNNT2, TNNI3, TNNCl, ACTC, MYBPC3
Z disk
TTN,
LBD3
VCL
CSRP3
TCAP
TTID
Nuclear
TAZ,
LMNA,
EMD,
EYA4,
SYNE1,
TMPO
RBM20
Mitochondrial
TTR, FRA,
Cox15, HFE
Cytoskeletal-Membrane Linkers
DMD, SGCG, SGCD, SGCB, SGCA, FKRP

Figure 4.1. Intracellular regions and the genes from these regions that cause DCM. From Dellefave & McNally (2010).

Clinical type	Inheritance	Gene name (symbol)
HDCM/DCM/RDCM	AD	Myosin heavy chain 7, (*MYH7*)
HDCM/DCM/atrial septal defect type 3	AD	Myosin heavy chain 6, (*MYH6*)
HDCM/DCM/RDCM/LVNC	AD	Troponin T2, cardiac (*TNNT2*)
HDCM/DCM/LVNC	AD	Tropomyosin 1 (*TPM1*)
HDCM/DCM	AD	Myosin binding protein 3, cardiac (*MYBPC3*)
HDCM/DCM/RDCM	AD, AR	Troponin I3, cardiac (*TNNI3*)
HDCM/DCM	AD	Actin alpha cardiac (*ACTC1*)
HDCM/DCM	AD/AR	Titin (*TTN*)
HDCM/DCM	AD	Troponin C1, cardiac (*TNNC1*)
HDCM/DCM	AD	Cystein- and glycine-rich protein 3 (*CSRP3*)
HDCM/DCM	AD	Titin cap (*TCAP*)
HDCM/DCM	AD	Vinculin (*VCL*)
HDCM/DCM	AD	Ankyrin repeat domain containing protein (*ANKRD1*)
DCM/RDCM	AD	Desmin (*DES*)
DCM	AD	Lamin A/C (*LMNA*)
DCM	AD	Sarcoglycan-delta (*SGCD*)
DCM	AD	Actinin alpha 2 (*ACTN2*)
DCM/LVNC	AD	Lim domain binding 3 (*LDB3*)
DCM	AD	Phospholamban (*PLB*)
DCM	AD	Presenilin 1 (*PSEN1*)
DCM	AD	Presenilin 2 (*PSEN2*)
DCM	AD	ATP binding cassette C9 (*ABCC9*)
DCM	AD	Sodium channel voltage-gated 5A (*SCN5A*)
DCM/HDCM	AD	Muscle-restricted coiled-coil (*MURC*)
DCM/HDCM	AD	Crystallin-alpha B (*CRYAB*)
DCM	AD	Four and a half Lim domains 2 (*FHL2*)
DCM	AD	Laminin alpha 4 (*LAMA4*)
DCM	AD	Nebulette (*NEBL*)
DCM/HDCM/RDCM	AD	Myopalladin (*MYPN*)
DCM	AD	RNA-binding motif protein 20 (*RBM20*)
HDCM/DCM	AD	Nexilin (*NEXN*)
DCM	AD	Bcl2-associated athanogene 3 (*BAG3*)
DCM	XR	Dystrophin (*DMD*)
DCM	XR	Emerin (*EMD*)
DCM/LVNC	XR	Tafazzin (*TAZ*)
DCM	XR	Fukutin (*FKTN*)
DCM/ARVC	AR	Desmoplakin (*DSP*)
DCM	AR	Dolichol kinase (*DOLK*)

Continuation of table 4.1

DCM	AR	GATA zinc finger domain-containing protein 1 (*GATAD1*)
DCM/ARVC	AR/AD	Plakoglobin (*JUP*)
DCM	AR	Flavoprotein (*SDHA*)

Table 4.1. DCM causative genes. The table is from Parvari *et al.* (2012). AD: autosomal dominant; AR: autosomal recessive; ARVC: arrhythmogenic right ventricular cardiomyopathy; DCM: dilated cardiomyopathy; HDCM: hypertrophic cardiomyopathy; RDCM: restrictive cardiomyopathy; LVNC: left ventricular noncompaction; XR-X-linked recessive.

Familial DCM risk is usually attributed to coding nucleotide variants that alter the amino acid sequence. For example, Tsubata *et al.* (2000) were able to identify a T451G change in the delta-sarcoglycan gene which was shared between affected individuals of a family suffering from DCM. This variant was not present in any of the unaffected individuals of the family, nor in a group of 200 controls. Mutations in sarcoglycans result in cytoskeletal abnormalities, and animal models with mutations in these genes frequently develop DCM (Towbin *et al.*, 1999). In another study, the exons of 10 genes, half of which comprised sarcomere genes, were sequenced in a group of 264 patients suffering from DCM (Lakdawala *et al.*, 2012). Forty clinically relevant variants were identified, none of which were present in a set of 200 healthy controls. This, together with the variants being found in genes known to cause DCM provides strong support for these variants influencing disease in these patients.

The family I analysed is depicted in figure 4.2. The pedigree shows that cases of DCM occur throughout the family, and both male and female individuals are affected. The pattern of DCM in this pedigree is consistent with an autosomal dominant mode of inheritance with incomplete penetrance. Sequence data was available for three cousins (v-5, v-13, v-15) and an uncle (iv-6), all four of whom suffer from DCM.

With regard to clinical presentation of the affected individuals, family member IV-6 had a heart transplant for DCM at the age of 54. Patient V-5 was diagnosed with DCM at ~40 years of age, and later died (although the age and cause of death are unknown). Patient V-15 was diagnosed with DCM at ~28 years of age, and with diabetes at ~37 years of age. Patient V-15 also displayed renal problems, however these were secondary to an renal tract outflow obstruction in infancy and thought to be unrelated to DCM. Family screening was triggered by the occurrence of multiple DCM cases in the pedigree and during the course of this process patient V-13, who was asymptomatic, was shown on echocardiography to have an established global dilated cardiomyopathy.

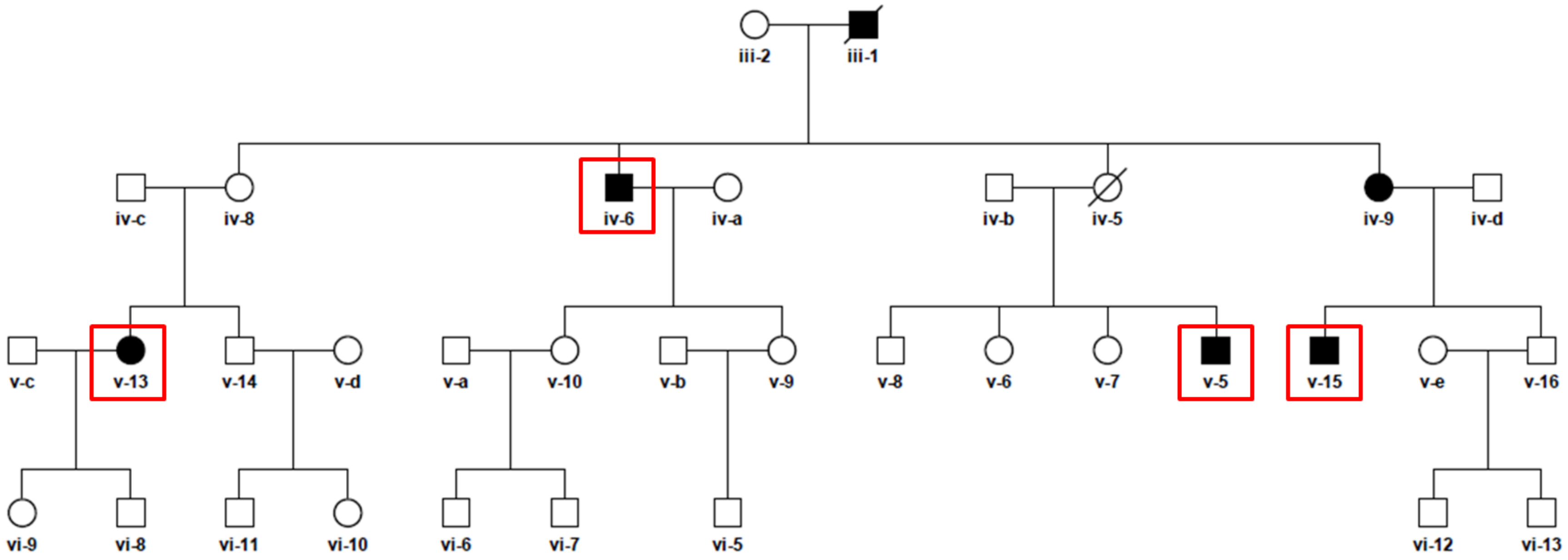


Figure 4.2. DCM pedigree. Sequence data was available for individual samples v-5, v-13, v-15, IV-6 (Indicated by a red squares).

4.2.2 Atrioventricular septal defects

AVSD is a complex disorder, covering a range of congenital heart conditions of varying severities (Hartman *et al.*, 2011). They involve incomplete septation of the atrioventricular valves and septa (Marino and Digilio, 2000; Craig, 2006; Miller *et al.*, 2010). The most severe form of AVSD, complete AVSD, occurs when both the atrial and ventricular septa do not develop properly, and when a single, common atrioventricular valve remains after development, as seen in figure 4.3A (Marino and Digilio, 2000; Robinson *et al.*, 2003). In the less severe form, partial AVSD, there is a deficiency of the atrial septum where there are separate right and left atrioventricular openings that do not close, figure 4.3B (Omeri *et al.*, 1965; Robinson *et al.*, 2003; Craig, 2006). Although complete AVSDs are detected at birth, children with the less severe forms may be asymptomatic so that detection occurs only at an older age (Robinson *et al.*, 2003; Minich *et al.*, 2010).

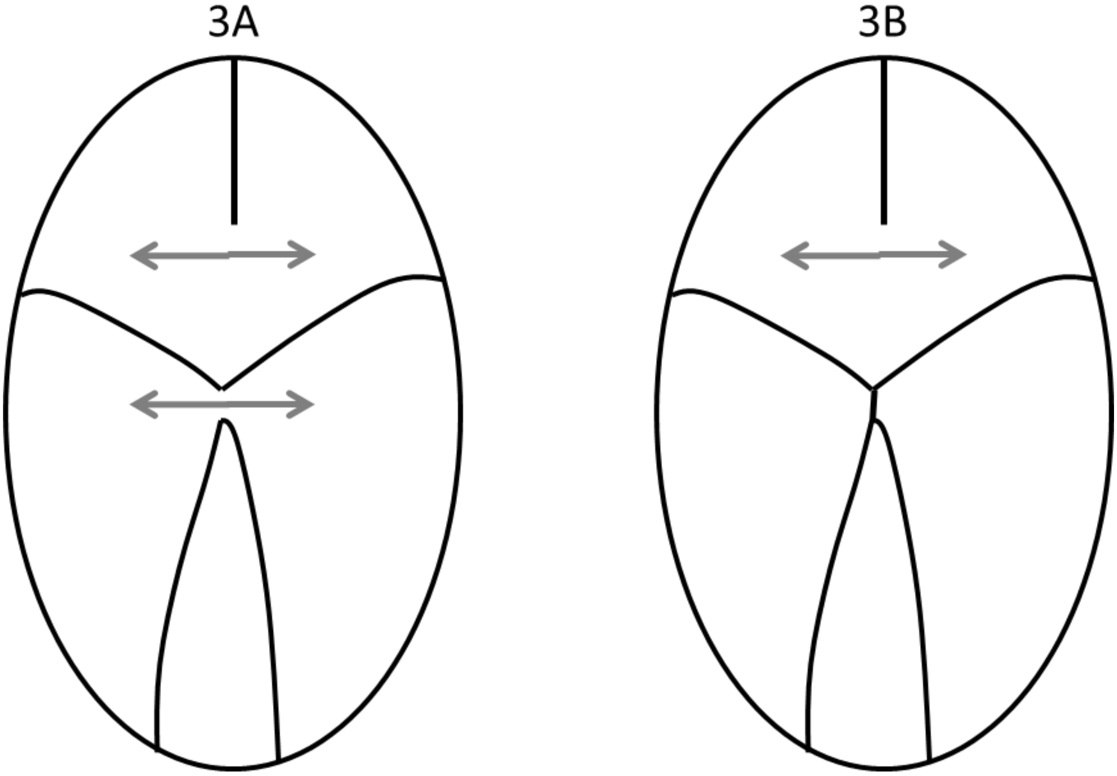


Figure 4.3 (A,B). Diagrammatic representation of AVSD. 4.3A = Complete AVSD. 4.3B = Partial AVSD.

AVSD’s account for ~5-7% of all congenital heart defects with an incidence of ~0.24-0.31 per 1000 births per year, and a prevalence of ~3.5-4.1 per 10 000 people (Craig, 2006; Reller *et al.*, 2008; Hartman *et al.*, 2011). They can occur as part of other recognised syndromes or as a single identifiable condition (Marino and Digilio, 2000; Robinson *et al.*, 2003; Hartman *et al.*, 2011). However, about 85% of cases are associated with other syndromes (Marino and Digilio, 2000). For example, AVSD is common in children with Down’s syndrome, with the complete form affecting between 17 - 50% of them (Marino and Digilio, 2000; Craig, 2006). Additionally, AVSD has been shown to be associated with the 3p-syndrome, a rare disease resulting from a deletion in chromosome 3, where it affects about a third of all cases (Green *et al.*, 2000).

As well as occurring as part of other identifiable syndromes (Sheffield *et al.*, 1997; Robinson *et al.*, 2003), there are some reported families in which multiple affected individuals occur with no other syndromic features (Wilson *et al.*, 1993; Kumar *et al.*, 1994; Sheffield *et al.*, 1997). This indicates a pattern consistent with autosomal dominant inheritance and incomplete penetrance (Marino and Digilio, 2000; Robinson *et al.*, 2003; Craig, 2006) in those families.

Various studies have been conducted in an attempt to identify the genes responsible for this disease. For example, a linkage study was performed on a large family containing 14 affected individuals (Sheffield *et al.*, 1997). Although the study failed to identify the particular genes responsible, a linkage region on chromosome 1 was identified. A further study by Robinson *et al.* (2003) analysed 50 unrelated individuals displaying full or partial AVSD to test for an association between *CRELD1* mutations and disease. Ten coding exons of *CRELD1* were sequenced in all the cases and they were able to identify three single base variants which were not identified in at least 100 controls and concluded that these represent disease associated variants. Also, the strong association of AVSD with Down Syndrome indicates causative genes on chromosome 21 (Locke *et al.*, 2010), and Trisomy 18 has also been implicated in some AVSD cases (Digilio *et al.*, 1999).

The pedigree I analysed is shown in figure 4.4 and contains 31 individuals, 10 of whom have AVSD. Sequence data was available for four of the family members (samples ii-5, iii-22, iv-30, iv-36). The occurrence of AVSD in the pedigree is consistent with an autosomal dominant mode of inheritance with incomplete penetrance (Wilson *et al.*, 1993). As AVSD’s occur predominantly as part of other syndromes, non-syndromic

Mendelian families such as these provide a very useful means of identifying possible causative genes and in understanding the disease further. The clinical findings for each member of the family are described in detail in the original paper by Wilson *et al.* (1993). Below, I will provide a brief description of the individuals analysed in the present study (As in Wilson *et al.* 1993).

Individual iv-30 was born with a complete AVSD with an aortic shelf coarctation. The aortic defect was treated using balloon dilation, and the AVSD surgically repaired. Facial features suggested the patient did not suffer from Down syndrome, and no other chromosomal abnormalities could be identified using high-resolution chromosome analysis. Although displaying no clinical evidence of AVSD, due to family history, individual ii-5 was described as an obligate carrier. This individual had three sons, none of which displayed any clinical symptoms of AVSD, one of which (iii-18), however, had a further two fully affected children. Individual iii-22, the son of obligate carrier ii-8, was affected and had to undergo surgery to repair the AVSD. Finally, individual iv-36, the son of obligate carrier ii-18, was born with AVSD, which was also repaired. This individual also had two brothers, one affected and one unaffected.

Wilson *et al.* (1993) performed a linkage study on this pedigree using a series of microsatellite polymorphisms along chromosome 21. Due to AVSDs being the predominant heart defect in children with Down Syndrome, they focused on the region of trisomy 21, and were able to exclude loci from this region as the cause of AVSD in this family.

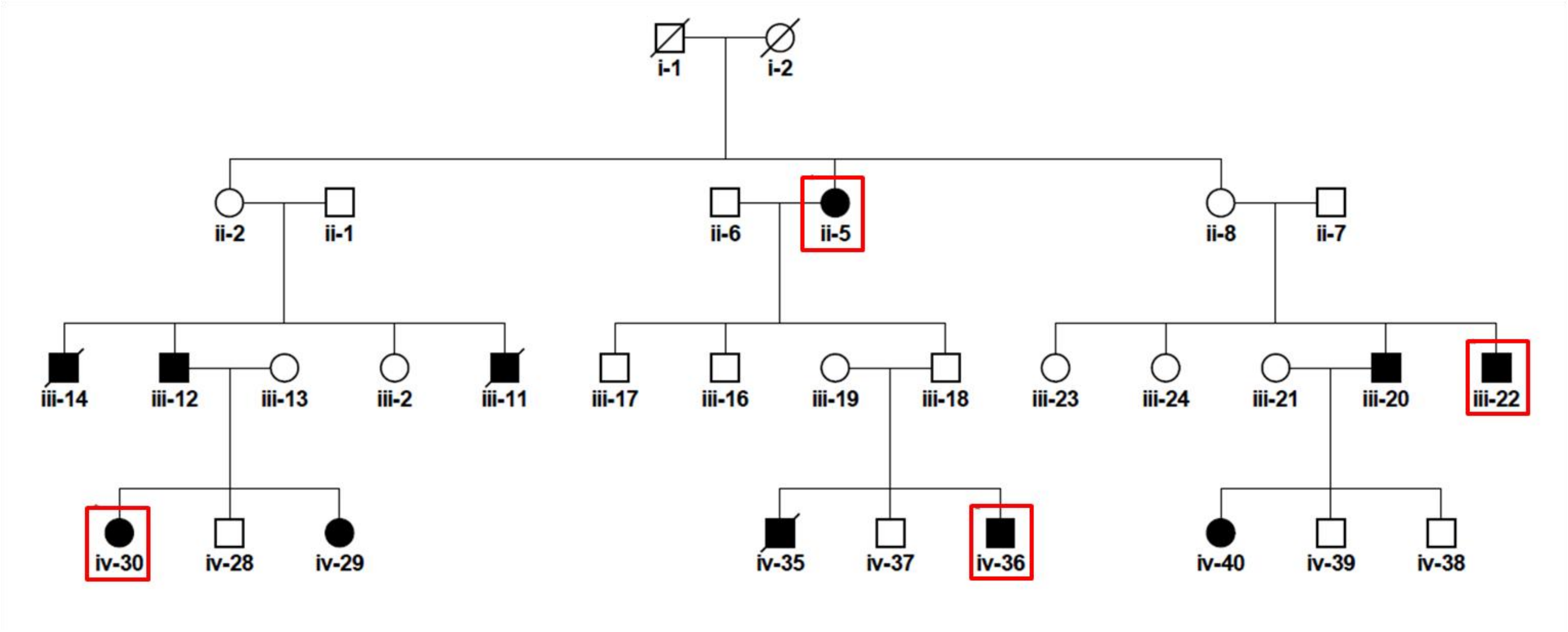


Figure 4.4. AVSD pedigree. Sequence data was available for individuals samples ii-5, iii-22, iv-30, iv-36 (Indicated by red squares).

4.2.3 Hereditary sclerosing poikiloderma

Hereditary sclerosing poikiloderma (HSP) is a very rare disease, and to my knowledge only 12 other cases have been described in the literature, (Lee *et al.*, 2012). In addition, I am aware of one other case from collaborators in Nantes, France, under Dr. Sébastien Küry, DVM at the Institut de Biologie (Service de Génétique Médicale, Laboratoire de génétique moléculaire). Although this case has not yet been published, Dr. Küry did provide sequence data for the unaffected parents and the affected offspring, which I analysed in addition to the sequence data generated in-house on the South African family described below.

HSP was first described in seven individuals from two unrelated families as a hereditary disorder displaying distinct features (Weary *et al.*, 1969) including widespread poikiloderma (A skin condition which can present with hyper/hypo-pigmentation and sclerosis), sclerosis of the palms and soles, linear or reticular hyperkeratotic and sclerotic bands in various regions, clubbing of the fingers, and calcinosis of tissues (Weary *et al.*, 1969; Lee *et al.*, 2012). Lee *et al.* (2012) describe the widespread poikiloderma and sclerotic bands as being the most important features in diagnosing the disease. As symptoms are not present at birth, detection only occurs in early childhood with the onset of progressive poikiloderma (Grau Salvat *et al.*, 1999).

Studies have noted that individuals affected with HSP and related family members sometimes also display cardiovascular abnormalities, such as heart valve defects and ventricular hypertrophy, thus also implicating cardiovascular disease as part of the phenotypic spectrum in HSP (Weary *et al.*, 1969; Grau Salvat *et al.*, 1999; Lee *et al.*, 2012). Most affected families indicate that HSP is consistent with an autosomal dominant form of inheritance with incomplete penetrance (Weary *et al.*, 1969; Khumalo *et al.*, 2006; Lee *et al.*, 2012). This appears to be the case in the pedigree I analysed, figure 4.5. The causative genetic defect for HSP has not yet been identified.

Five members of the pedigree I analysed suffer from HSP. A full clinical report on each patient is given in Khumalo et al. (2006). A description for the three family members for which sequence data were available follows. The family members included two affected siblings (ii-4 and ii-5) and an unaffected mother (i-3). During childhood individual ii-5 displayed heat intolerance and skin lesions including hyper- and hypo-pigmentation and epidermal atrophy, and by age 9 had developed Achilles tendon contractures. The tendon contractures were surgically treated at age 14. By adulthood she had also developed telangiectasias, a mottled skin, epidermal atrophy around the face, fine to no body hair, and tendon abnormalities. Her arms and legs had virtually no hair and mottled skin pigmentation. She had atrophy of both thenar and hypothenar eminences and was unable to fully extend her fingers. She also suffered from hypohidrosis, which together with the sparse hair, suggested ectodermal dysplasia. Her brother, individual ii-4, displayed similar skin and limb characteristics, as well as heat intolerance, but suffers from no tendon abnormalities (Khumalo *et al.,* 2006). He suffered a slight cardiomegaly and left ventricular hypertrophy.

Two other affected patients, the father (i-2) and an a half brother (ii-3), were said to have suffered from similar conditions. However, both had died as a result of pulmonary fibrosis prior to the present study commencing and no sequence data were available for them (Khumalo *et al*., 2006). However, collaborators in South Africa were able to extract DNA from the deceased brother using a paraffin embedded liver sample.

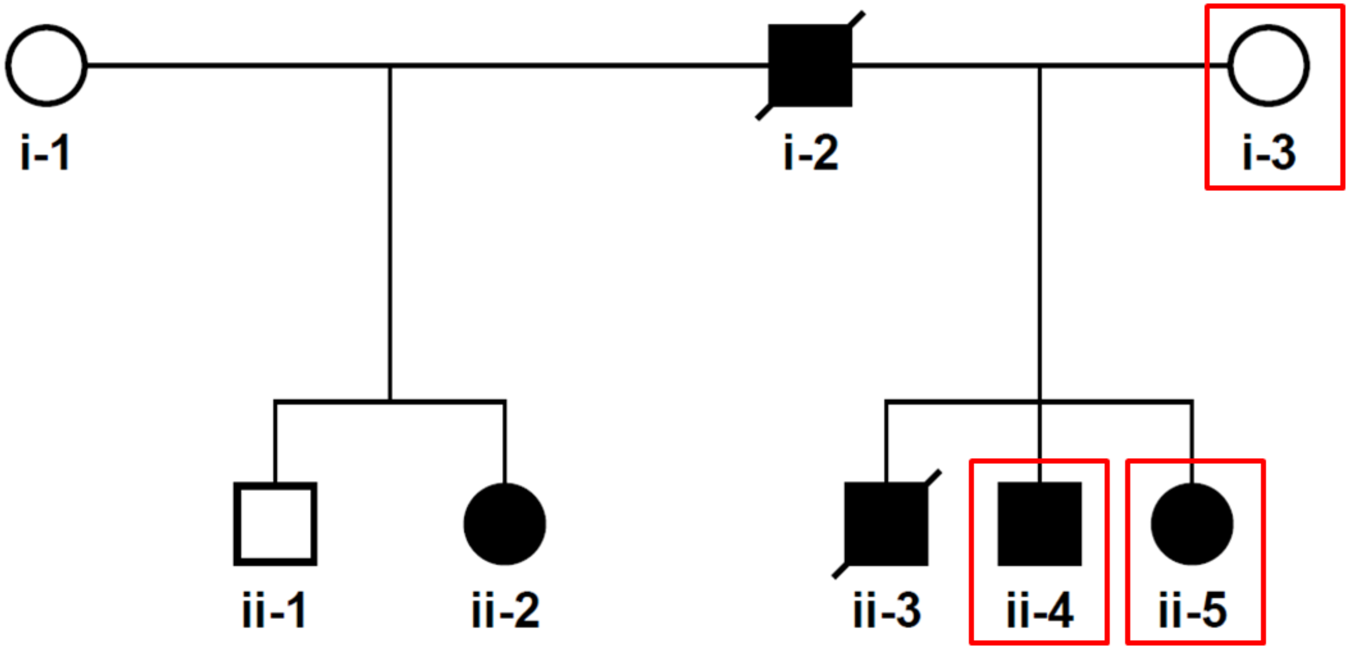


Figure 4.5. HSP pedigree. Sequence data was available for individuals samples ii-4, ii-5, and i-3 (Indicated by red squares).

4.3 Materials and Methods

4.3.1 Samples and sequencing

DNA was extracted from blood using standard methods. This was done in the laboratory at The University of Newcastle by Thahira Rahman and Rafiqul Hussain, and described in chapter 2. Targeted NGS sequencing was performed as described in chapter 2. In the current study, target enrichment was carried out using the SureSelect Agilent 38Mb Human All Exon targeting kit (http://www.genomics.agilent.com).

4.3.2 Sequence analysis

GERALD (CASAVA-1.6.0) was used as base caller and to generate the Fastq files. Read quality was assessed using average base quality, calculated using the NGS QC Toolkit v2.2.3 (Patel and Jain, 2012).

The NovoAlign aligner v2.07.13 (http://www.novocraft.com) was used to map the sequences to the Human Genome (Build 37) and the Samtools/bcftools v0.1.17 (Li *et al.*, 2009) package was originally used to identify both single base substitution and indel variants. However, it was later decided that the BWA-Dindel pipeline would be a better pipeline for identifying indels (see chapter 3). Therefore, BWA-Dindel was also used as a method to identify indels. I will present indel results from both pipelines. Aside from the selection of only unique alignments in NovoAlign, default alignment and variant calling criteria were used in all programmes.

Variants were filtered and annotated using custom written perl scripts. MutationTaster v20100416 (Schwarz *et al.*, 2010) and wAnnovar v2011-11-20 (Wang *et al.*, 2010b), were used to annotate variants and to assess potential variant pathogenicity.

4.3.3 Variant filtering

There is a risk that variant filtering methods could inadvertently remove the disease causing variant. Therefore, in cases where a potentially disease causing variant in a plausible candidate gene of interest was not identified I implemented an additional, less stringent set of filtering criteria, filtering set B (Figure 4.6). Both sets are described in detail below.

The first step of filtering set A (Stringent filtering) involved the selection of only the on-target variants using the Agilent Exome target positions. Indel variants were allowed to lie within 500 base pairs of the target positions, while single base substitutions had to be completely located within the target regions. All homozygous variants were then removed and only those variants present in all of the affected individuals were selected. Variants occurring within the Exome Variant Server (EVS) and 1000 Genomes databases at a frequency exceeding 1% were then also removed. A control list of variants was also used as a filter to remove possible errors. The control list contains single base substitutions from 119 unrelated exomes, and indel variants from 114 unrelated exomes. All exome data comprising the control lists had been sequenced in-house using the Agilent 50Mb Whole Exome Targeting kit and the Illumina Genome Analyser IIx. The list was compiled by Dr. Helen Griffin (2012, *pers. comm*). All variants present in the control list at a frequency of more than 1% were removed. Only coding, non-synonymous and splice variants were then selected for analysis with MutationTaster.

The first step of filtering set B (Less stringent) involved the selection of only the on-target variants from the Agilent Exome target positions. Indel variants were allowed to lie within 500 base pairs of the target positions, while single base substitutions had to be completely located within the bait regions. The second filtering step involved selecting only heterozygous variants present in all of the affected individuals. All variants occurring in the EVS and 1000 Genomes databases were then removed, as well as those occurring within the control list, at a frequency exceeding 1%. As well as coding, non-synonymous and splice variants, filtering set B also included non-coding and synonymous variants for assessment using MutationTaster, making it less stringent than filtering set A (See figure 4.7).

4.3.4 Variant validations

Variants were validated using Sanger sequencing, in labs at The University of Newcastle by Dr. Elise Glen.

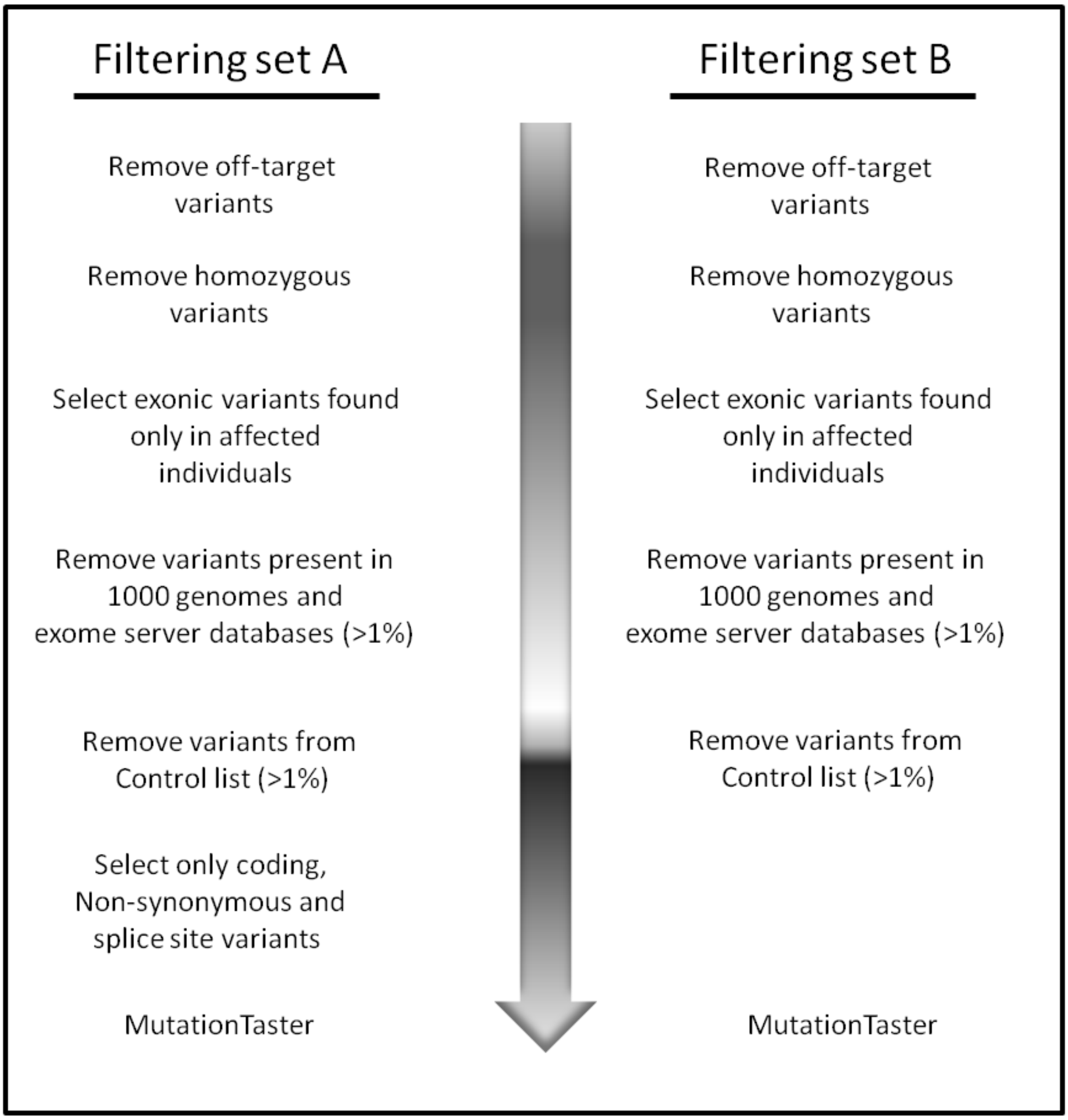

Filtering set A
Filtering set B
Remove off-target variants
Remove homozygous variants
Select exonic variants found only in affected individuals
Remove variants present in 1000 genomes and exome server databases (>1%)
Remove variants from Control list (>1%)
Select only coding, Non-synonymous and splice site variants
MutationTaster
Remove off-target variants
Remove homozygous variants
Select exonic variants found only in affected individuals
Remove variants present in 1000 genomes and exome server databases (>1%)
Remove variants from Control list (>1%)
MutationTaster

Figure 4.6. Filtering steps adopted in both filtering set A and filtering Set B.

4.4 Results

The results section is divided into two parts. The first provides an overview of the sequencing and variant call results, including details on the amount of sequence data that was generated, the sequence and alignment quality, sensitivity and specificity estimates, and the effect of the filtering criteria. The second part provides more detailed results on the potentially disease causing variants that were identified in each family.

4.4.1 Sequence and variant call overview

Between ~107x10^6 and ~160x10^6 paired end reads were generated across all of the samples. A mean target base coverage of between 102 and 191 was achieved across all samples and over 97% of the target bases were covered at least 1 fold, and over 85% were covered at least 20 fold (Table 4.2). Where I looked, the base quality was above 35 for at least the first 40 bases of each read, only dropping down to ~30 towards the ends of the reads.

Samples	Mean target coverage	%bases > 20fold	%bases > 10fold	%bases > 5fold	%bases > 1fold
DCM Patient 1 - Affected	177.63	90.03	94.3	96.46	98.37
DCM Patient 2 - Affected	177.63	86.73	91.84	94.83	97.99
DCM Patient 3 - Affected	177.63	86.41	91.48	94.52	97.91
DCM Patient 4 - Affected	177.63	87.07	92.8	95.74	98.28
AVSD Patient 1 - Affected	121.47	85.29	91.37	94.61	97.3
AVSD Patient 2 - Affected	116.83	86.54	92.63	95.65	98.17
AVSD Patient 3 - Affected	144.07	86.19	92.19	95.41	98.18
AVSD Patient 4 - Affected	150.89	90.37	94.61	96.64	98.29
HSP Patient 1 - Affected	191.62	86.95	91.98	94.89	97.73
HSP Patient 2 - Affected	177.63	90.04	94.3	96.46	98.37
HSP Patient 3 - Unaffected	179.88	88.66	93.44	95.9	98.15

Table 4.2. Sample summary statistics. The table describes the mean target base coverage and percentage target bases covered 20 fold and above, 10 fold and above, 5 fold and above, and 1 fold and above. Values for each patient of each pedigree are given.

More than 93% of the variants could be removed by selecting only the on-target variants, those variants shared amongst affected individuals, and by removing the homozygous changes (Appendix figures 4.1 – 4.3). Therefore, a very large proportion of the variants could be removed by using only the first three filtering steps.

4.4.2 Dilated cardiomyopathy family

In excess of 200000 variants were identified across all the samples from this family, of which at least 88% comprised single base substitutions (Figures 4.7). After applying filtering set A, 4 variants remained (Table 4.3). Only the *HYDIN* variant was predicted as disease causing by MutationTaster. However, this variant was also observed in all samples from the AVSD and the HSP pedigrees, and is therefore not a good candidate for disease in this family.

The filtering criteria were then relaxed (Filtering set B), after which 11 variants remained (Table 4.3). Three were predicted as disease causing by MutationTaster, the *HYDIN* mutation, the *SLC38A10* mutation, and a splice site mutation in the *ANKRD20A1* (chr9, 67927076). Out of these three only the *ANKRD20A1* variant is not a recognised SNP, and *ANKRD20A1* is related to *ANKRD1* which can cause DCM.

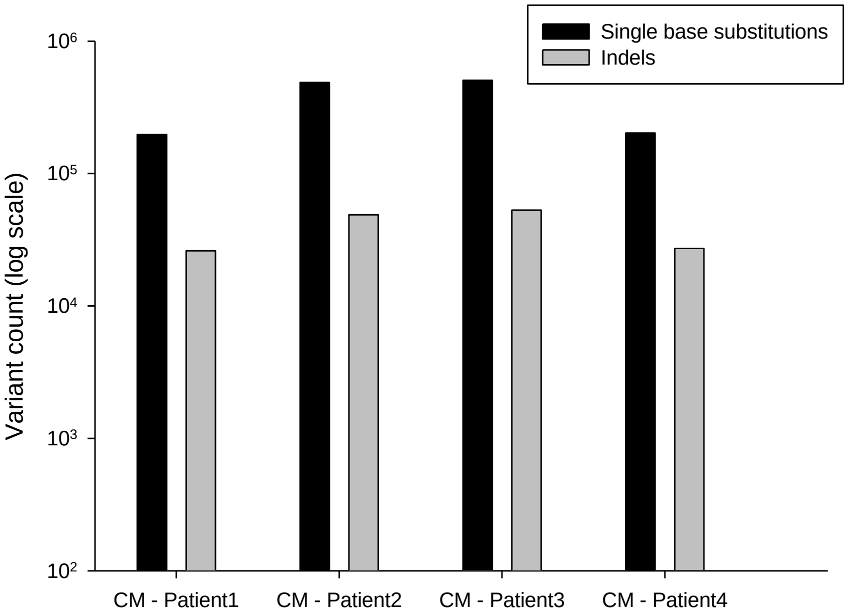


Figure 4.7. Number of unfiltered single base substitutions and indels identified using NovoAlign-Samtools for all patients in the DCM pedigree.

Chrom.	Position	Reference	Indel	Gene	EVS	1000G	dbSNP135	Control list	Filter set
chr1	87045902	ACCTAC	-	*CLCA4*	0	0	rs77067122	0	B
chr9	38397083	G	A	*ALDH1B1*	0.005856	0.0023	rs41278335	0	B
chr9	67927076	G	A	*ANKRD20A1*	0	0		0	B
chr9	67968476	C	T	*ANKRD20A1, ANKRD20A3*	0	0	rs4055530	0	A,B
chr10	46999607	-	AGGTGGGGG	*GPRIN2*	0	0	rs58801928	0	B
chr10	126463282	T	C	*METTL10*	0.000197	0	rs139315006	0	A,B
chr11	17352482	CAA	-	*NUCB2*	0	0	rs72423941	0	B
chr11	56467881	T	C	*OR9G1, OR9G9*	0	0	rs73474900	0	B
chr11	56468212	G	A	*OR9G1, OR9G9*	0	0	rs591369	0	A,B
chr16	70896017	A	-	*HYDIN*	0	0	rs57797337	0	A, B
chr17	79219505	TGA	-	*SLC38A10*	0	0	rs3833102	0	B

Table 4.3. Variants shared between all affecteds which passed the filtering steps in the DCM pedigree. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

Using the BWA-Dindel pipeline for indel identification, I was unable to identify any indels which passed either filtering set A (Stringent) or filtering set B (Less stringent).

4.4.3 Atrioventricular septal defect family

In excess of 170000 variants were identified across all the samples of this family, of which at least 88% were single base substitutions (Figure 4.8). After stringent filtering (Filtering set A), only the *HYDIN* variant was described by MutationTaster as potentially disease causing (Table 4.4). However, this variant was also identified in all the samples from both the DCM and HSP pedigree.

Therefore, the relaxed set of filtering criteria was applied to the data (Filtering set B), however this resulted in only one additional variant being identified in the *OR9G1* gene (Table 4.4), which is a synonymous change and not predicted as disease causing by MutationTaster.

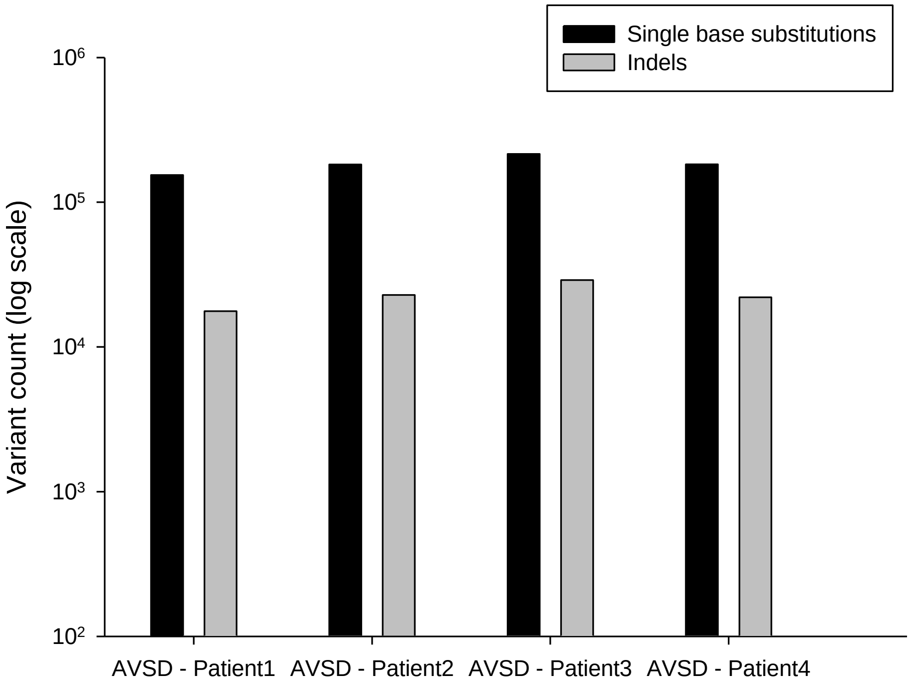


Figure 4.8. Number of unfiltered single base substitutions and indels identified using NovoAlign-Samtools for all patients in the AVSD pedigree.

Chr.	Position	Reference	Indel	Gene	EVS	1000G	dbSNP135	Control list	Filter set
chr6	30558478	-	A	*ABCF1*	0	0	rs4148252	0	A,B
chr11	56467881	T	C	*OR9G1,OR9G9*	0	0	rs73474900	0	B
chr11	56468212	G	A	*OR9G1,OR9G9*	0	0	rs591369	0	A,B
chr16	70896017	A	-	*HYDIN*	0	0	rs57797337	0	A,B

Table 4.4. Variants shared between all affecteds which passed the less stringent filtering steps in the AVSD pedigree. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

Using the BWA-Dindel pipeline, 5 indels were predicted as disease causing (Table 4.5). The same 5 were identified and predicted as disease causing using both filtering set A (Stringent), and filtering set B (Less stringent). None were in plausible candidate genes.

Chromosome	**Position**	**Reference**	**Indel**	**Gene**	**EVS**	**1000G**	**dbSNP135**	**Control list**
chr6	29912029	G	-	*HLA-A*	0	0	rs149455102	0
chr16	70896016	A	-	*HYDIN*	0	0	rs11337008	0
chr16	81242149	TT	-	*PKD1L2*	0	0	rs150289691	0
chr17	21319651	GAG	-	*KCNJ12*	0	0	rs112163749	0
chr22	38120176	CCT	-	*TRIOBP*	0	0	rs146565844	0

Table 4.5. Variants shared between all affecteds identified using the BWA-Dindel pipeline and which passed both filtering sets. All were predicted as disease causing by MutationTaster. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

4.4.4 Hereditary Sclerosing Poikioderma family

In excess of 88000 variants were identified for all of the samples from this family, comprising at least 88% single base substitutions (Figure 4.9). Using the stringent set of filtering criteria (Filtering set A) I was able to identify 56 variants (Table 4.6), of these 30 were predicted as potentially disease causing by MutationTaster (Table 4.7). Of particular interest are the variants in the *BLK* and *ALOXE3* genes, as previous studies indicate that changes to these two genes result in skin disorders presenting with keratosis and icthyosis (Starfield *et al.*, 1997; Appel *et al.*, 2002; Jobard *et al.*, 2002). However, both have been observed in the EVS and 1000 Genomes data, substantially reducing their candidacy for causing HSP.

Collaborators in Nantes provided exome sequencing data on a trio family consisting of one affected offspring with HSP with a phenotype very similar to that described by Khumalo *et al.* 2006 (the family sequenced in-house). In this family, which was simplex, a strategy involving searching for *de novo* variants in the affected offspring was followed. This process identified a novel, non-synonymous variant in the

FAM111B gene (c.1789A>G) which is unreported in the dbSNP, 1000 Genomes or EVS databases. I also identified a non-synonymous variant in this gene (c.1771T>G; Table 4.6), which was present in both affected siblings, but not in the unaffected mother. The variant I identified within the *FAM111B* gene has been validated, alters the amino acid, and is not located in either the dbSNP135 or 1000 genomes databases. Also *FAM111B* was validated in the affected, deceased brother using DNA extracted from a paraffin embedded liver sample.

I reanalysed their sequencing data using my pipeline and after stringent filtering were able to identify 101 *de novo* variants, including the *FAM111B* variant (Table 4.8). In addition to the *FAM111B*, two additional genes were shared between the two pedigrees (Table 4.9). Of these, the variants in the *CNTNAP3B* present in my sample and my collaborators' patient are known SNPs, rs62558062 and rs3739621 respectively. However, these are possibly very rare SNPs as they are not present in the EVS, or 1000 genomes databases, or in the control list.

FAM111B is a gene of unknown function which to date has not been implicated in any human disease. Expression analysis, performed by Dr. Elise Glen, showed that the gene is expressed in skin fibroblasts.

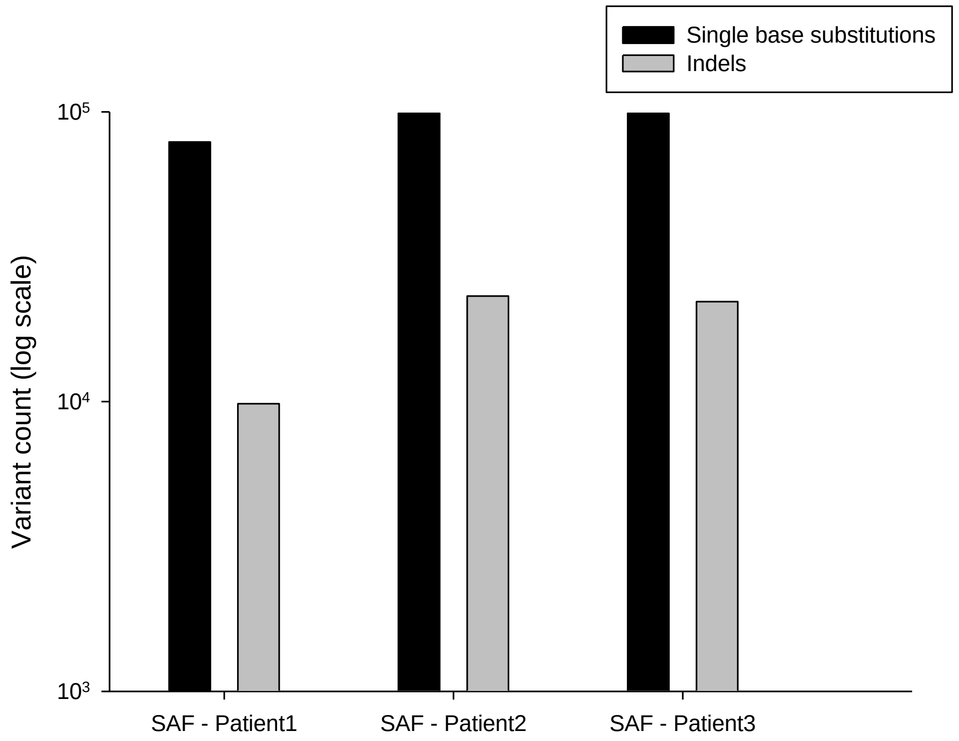


Figure 4.9. Unfiltered single base substitutions and indels identified using NovoAlign-Samtools for all patients in the HSP pedigree.

Chr.	**Position**	**Reference**	**Indel**	**Gene**	**EVS**	**1000G**	**dbSNP135**	**Control list**
chr2	1457549	G	C	*TPO*	0	0		0
chr2	17692189	C	T	*RAD51AP2*	0.001258	0.0009	rs183882477	0
chr2	24929631	A	G	*NCOA1*	0	0		0
chr2	242169660	C	T	*HDLBP*	0	0		0
chr3	19479731	G	A	*KCNH8*	0.001208	0.0014	rs138531032	0
chr3	45942584	C	T	*CCR9*	0.001301	0.0009	rs139107036	0.003521127
chr4	8416586	T	G	*ACOX3*	0.007436	0.01	rs73211315	0.007042254
chr5	56778305	A	G	*ACTBL2*	0.003346	0.0005	rs148214432	0.003521127
chr5	169446042	G	A	*DOCK2*	0.000186	0	rs149008494	0
chr6	90422360	C	T	*MDN1*	0.003718	0.0009	rs62417304	0.003521127
chr6	126210797	G	A	*NCOA7*	0.004467	0.0018	rs35223550	0
chr7	26224760	G	A	*NFE2L3*	0.003811	0.0037	rs148159120	0.007042254
chr7	120965470	-	CCCA	*WNT16*	0	0	rs55710688	0
chr8	11412934	G	A	*BLK*	0.003625	0.0009	rs141865425	0.007042254
chr8	20036702	C	T	*SLC18A1*	0.001115	0	rs17215808	0
chr8	30916058	A	G	*WRN*	0.00316	0.0014	rs34477820	0.003521127
chr8	33361016	C	T	*TTI2*	0.001208	0	rs150984360	0
chr9	20995555	C	T	*FOCAD*	0.001859	0.0014	rs145021526	0
chr9	33953294	G	A	*UBAP2*	0.001022	0.0009	rs150275904	0.003521127
chr9	43685298	G	T	*CNTNAP3B*	0	0	rs62558062	0
chr9	125391777	-	A	*OR1B1*	0	0	rs11421222	0
chr10	50732139	C	T	*ERCC6*	0.009481	0.01	rs4253047	0.003521127
chr10	51768664	G	T	*AGAP6*	0	0		0
chr11	4389407	G	-	*OR52B4*	0	0	rs11310407	0
chr11	58893431	T	G	*FAM111B*	0	0		0
chr11	62381864	G	A	*ROM1*	0	0		0
chr11	63487475	G	C	*RTN3*	0.0066	0.0023	rs7936660	0.003521127
chr11	66360021	C	T	*CCDC87*	0.007157	0.0046	rs1110707	0.003521127
chr11	66468736	C	T	*SPTBN2*	0.000093	0		0
chr12	4737404	C	T	*AKAP3*	0.005298	0.0018	rs71579261	0
chr13	75887003	T	C	*TBC1D4*	0.00197	0.01	rs149821147	0
chr13	96511850	A	G	*UGGT2*	0.007639	0.01	rs9525072	0.007042254
chr16	14028081	C	T	*ERCC4*	0.004276	0.0032	rs1799802	0
chr16	20335264	C	T	*GP2*	0.003253	0.0032	rs145297751	0
chr16	55844871	T	C	*CES1*	0.001394	0	rs140704082	0.003521127
chr16	56871605	G	A	*NUP93*	0	0		0
chr16	67000764	C	T	*CES3*	0	0		0
chr16	71318184	C	G	*FTSJD1*	0	0		0
chr16	85697023	G	A	*KIAA0182*	0.000651	0	rs146762745	0

Table 4.6 Continued

chr17	3594281	G	-	*P2RX5*	0	0	rs3215407	0
chr17	7224519	T	A	*NEURL4*	0.005058	0.0037	rs145900596	0
chr17	8006708	G	A	*ALOXE3*	0.000558	0.0009	rs147149459	0
chr17	8195873	C	T	*SLC25A35*	0.000279	0	rs146737646	0
chr17	10322333	C	G	*MYH8*	0.003067	0	rs146732664	0
chr17	10409243	G	A	*MYH1*	0.003067	0.0005	rs142560385	0
chr17	11572991	T	G	*DNAH9*	0.000093	0	rs142009409	0
chr17	46620525	G	C	*HOXB2*	0.000372	0		0
chr18	12329644	G	A	*AFG3L2*	0.001208	0.0018	rs117182113	0
chr18	23619302	A	T	*SS18*	0	0		0
chr19	6754659	A	T	*SH2D3A*	0	0		0
chr19	21132126	G	A	*ZNF85*	0.00084	0.0014	rs140775014	0
chr19	21477325	A	T	*ZNF708*	0.002231	0.0014	rs77583547	0
chr19	52004794	-	C	*SIGLEC12*	0	0	rs67024588	0
chr19	52004795	-	T	*SIGLEC12*	0	0		0
chr22	41616779	G	A	*L3MBTL2*	0	0		0
chr22	45198034	C	T	*ARHGAP8, PRR5-ARHGAP8*	0.006507	0.01	rs41278883	0

Table 4.6. Variants which passed the stringent filtering steps in the HSP pedigree. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

Chr.	Position	Reference	Indel	Gene
chr1	150199051	TTCCTC	-	*ANP32E*
chr2	242169660	C	T	*HDLBP*
chr3	19479731	G	A	*KCNH8*
chr3	45942584	C	T	*CCR9*
chr5	56778305	A	G	*ACTBL2*
chr6	126210797	G	A	*NCOA7*
chr7	26224760	G	A	*NFE2L3*
chr8	11412934	G	A	*BLK*
chr8	20036702	C	T	*SLC18A1*
chr9	20995555	C	T	*KIAA1797*
chr9	33953294	G	A	*UBAP2*
chr9	125391777	-	A	*OR1B1*
chr10	51768664	G	T	*AGAP6*
chr11	4389407	G	-	*OR52B4*
chr13	75887003	T	C	*TBC1D4*
chr13	96511850	A	G	*UGGT2*
chr15	71276488	AAC	-	*LRRC49*
chr16	10524660	GAC	-	*ATF7IP2*
chr16	14028081	C	T	*ERCC4*
chr16	56871605	G	A	*NUP93*
chr16	71318184	C	G	*FTSJD1*
chr17	3594281	G	-	*P2RX5*
chr17	8006708	G	A	*ALOXE3*
chr17	10409243	G	A	*MYH1*
chr17	46620525	G	C	*HOXB2*
chr18	12329644	G	A	*AFG3L2*
chr18	23619302	A	T	*SS18*
chr19	30500143	TGA	-	*URI1*
chr22	41616779	G	A	*L3MBTL2*
chr22	45198034	C	T	*PRR5-ARHGAP8*

Table 4.7. Variants which passed the strict filtering set and were identified as potentially disease causing by MutationTaster in the HSP pedigree. All positions based on HG19 reference.

Chr.	Position	Reference	Indel	Gene	EVS	1000G	dbSNP135	Control list
chr1	2126139	C	G	*C1orf86*	0	0	rs6662296	0
chr1	2433578	C	A	*PLCH2*	0	0		0
chr1	108152557	G	T	*VAV3*	0.001766	0.0014	rs138334746	0.002092
chr1	117158857	C	T	*IGSF3*	0	0		0.002092
chr1	12907408	T	A	*HNRNPCL1, LOC649330*	0.000651	0	rs146075045	0.008368
chr1	149902766	G	A	*MTMR11*	0	0		0
chr1	156438602	T	C	*MEF2D*	0	0		0
chr1	161514691	A	T	*FCGR3A*	0	0		0
chr1	230561391	C	A	*PGBD5*	0	0		0
chr1	247615264	G	-	*OR2B11*	0	0		0
chr2	74466662	A	G	*SLC4A5*	0	0		0
chr2	109098822	T	C	*GCC2*	0	0		0
chr2	236761415	-	GGGC	*AGAP1*	0	0		0
chr2	240029799	T	G	*HDAC4*	0	0		0
chr3	49329992	G	T	*USP4*	0	0		0
chr3	73673586	GC	-	*PDZRN3*	0	0		0
chr4	159590833	C	T	*C4orf46*	0	0		0
chr5	7820771	T	C	*ADCY2*	0	0		0
chr5	139422562	C	G	*NRG2*	0	0		0
chr5	140604659	G	A	*PCDHB14*	0	0		0
chr6	26444248	T	A	*BTN3A3*	0	0		0
chr6	29911119	G	C	*HLA-A*	0	0	rs3173419	0.006276
chr7	2353998	G	T	*SNX8*	0	0		0
chr7	2552898	-	GTGG	*LFNG*	0	0		0
chr7	48349604	C	G	*ABCA13*	0	0		0
chr7	51098567	GTCT	-	*COBL*	0	0		0
chr7	73249193	-	TTCCA CAGGCG	*WBSCR27*	0	0		0
chr7	73249197	-	TCAGG CGGTCC	*WBSCR27*	0	0		0
chr7	95926236	C	T	*SLC25A13*	0	0		0
chr7	149506211	-	G	*SSPO*	0	0		0
chr7	151684361	C	A	*GALNTL5*	0	0		0
chr7	153750014	G	A	*DPP6*	0	0	rs2240820	0
chr8	22436870	C	A	*PDLIM2*	0	0		0
chr8	25279148	G	A	*GNRH1*	0	0		0

Table 4.8 continued

chr8	52732981	C	G	*PDCMTD1*	0	0		0
chr9	21228151	G	C	*IFNA17*	0	0		0
chr9	33558121	G	T	*ANKRD18B*	0	0		0
chr9	33796703	C	G	*PRSS3*	0	0		0
chr9	43822704	G	A	*CNTNAP3B*	0	0	rs3739621	0
chr9	139964853	G	C	*SAPCD2*	0	0		0
chr10	51748530	-	C	*AGAP6*	0	0		0
chr10	51827896	C	T	*FAM21A*	0	0	rs11552619	0
chr10	74790045	G	A	*P4HA1*	0	0		0
chr10	81272467	A	T	*EIF5AL1*	0	0		0
chr10	121196274	G	T	*GRK5*	0	0		0
chr10	125780764	GT	GGGT	*CHST15*	0	0		0
chr10	126312137	C	T	*FAM53B*	0.000093	0		0
chr11	1078654	G	T	*MUC2*	0	0		0
chr11	32119977	C	A	*RCN1*	0	0		0
chr11	46342260	-	T	*CREB3L1*	0	0		0
chr11	49974777	C	T	*OR4C13*	0	0		0
chr11	58893449	A	G	*FAM111B*	0	0		0
chr11	64669850	C	G	*ATG2A*	0.000469	0	rs149707582	0
chr12	10167883	C	T	*CLEC12B*	0	0		0
chr12	52629122	C	T	*KRT7*	0	0		0
chr12	56094151	G	A	*ITGA7*	0	0		0
chr12	117977558	C	-	*KSR2*	0	0		0
chr12	132633381	T	C	*NOC4L*	0	0		0
chr13	28942761	G	C	*FLT1*	0	0		0
chr13	52650273	C	T	*NEK5*	0.000558	0	rs139136964	0.002092
chr13	78272278	-	C	*SLAIN1*	0	0	rs71102772	0
chr14	93176029	C	A	*LGMN*	0	0		0
chr15	31521516	T	-	*LOC283710*	0	0		0
chr15	35086927	G	A	*ACTC1*	0	0		0
chr15	40545052	G	A	*C15orf56*	0	0		0
chr15	75131978	C	T	*ULK3*	0.000207	0		0
chr15	75981901	C	T	*CSPG4*	0	0		0
chr16	2159179	G	A	*PKD1*	0	0		0
chr16	15112733	G	C	*PDXDC1*	0	0		0
chr16	15489840	C	A	*MPV17L*	0	0		0
chr16	88677735	G	T	*ZC3H18*	0	0		0
chr16	88772985	C	A	*CTU2*	0	0		0

Table 4.8 continued

chr17	7734052	C	T	*DNAH2*	0	0		0
chr17	34499245	G	C	*TBC1D3B*	0.0051	0		0.004184
chr17	39296135	G	A	*KRTAP4-6*	0	0.0037	rs28405099	0
chr17	40336172	TC	-	*HCRT*	0	0		0
chr17	44626083	C	A	*LRRC37A2*	0	0		0
chr17	61660896	T	-	*DCAF7*	0	0		0
chr18	72997837	A	C	*TSHZ1*	0	0		0
chr19	5610086	C	A	*SAFB2*	0	0		0
chr19	16582756	T	C	*EPS15L1*	0	0		0
chr19	33355209	T	C	*SLC7A9*	0	0		0
chr19	35504178	C	A	*GRAMD1A*	0	0		0
chr19	40421674	G	T	*FCGBP*	0	0		0
chr19	41060188	G	T	*SPTBN4*	0	0		0
chr19	43411160	G	C	*PSG6*	0.000093	0	rs140788501	0
chr19	50040423	C	A	*RCN3*	0.000279	0.0005	rs142564622	0
chr20	126310	AC	-	*DEFB126*	0	0		0
chr20	2083466	A	T	*STK35*	0	0		0
chr20	19261648	T	C	*SLC24A3*	0	0		0
chr20	23965998	T	G	*GGTLC1*	0	0	rs62195276	0
chr20	62065186	C	A	*KCNQ2*	0	0		0
chr21	36042747	G	T	*CLIC6*	0	0		0
chr22	41252508	C	T	*ST13*	0	0		0
chrX	8699935	C	T	*KAL1*	0	0		0
chrX	48895943	T	C	*TFE3*	0	0		0
chrX	54780125	T	A	*ITIH6*	0	0		0
chrX	100749038	C	T	*ARMCX4*	0	0	rs34379067	0
chrX	100749041	A	G	*ARMCX4*	0	0		0
chrX	111000833	C	G	*ALG13*	0	0		0
chrX	153690631	G	A	*PLXNA3*	0.008909	0.0036	rs141197316	0.006276

Table 4.8. *De novo* variants which passed the stringent filtering steps in the second HSP pedigree identified by my collaborators. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

Gene	Chr,position,reference,variant (Patient sample)	Chr,position,reference,variant (Collaborators patient)
AGAP6	chr10, 51768664,G,T	chr10,51748530, -,C
CNTNAP3B	chr9,43685298,G,T	chr9,43822704,G,A
FAM111B	chr11,58893431,T,G	chr11,58893449,A,G

Table 4.9. Genes in which variants were identified in both HSP pedigrees. All positions based on HG19 reference.

Using the BWA-Dindel pipeline as an additional indel calling pipeline, 4 indels passed filtering set A and were predicted as potentially disease causing, and 5 passed filtering set B and were predicted as potentially disease causing (Table 4.10). None of the variants were identified in plausible candidate genes of interest.

Chromosome	Position	Reference	Variant	Gene	EVS	1000G	dbSNP135	Control list	Filtering set
chr10	55582230	AGG	-	*PCDH15*	0	0	-	0	A,B
chr10	127668864	GAA	-	*FANK1*	0	0.01	rs146106149	0	A,B
chr19	30500119	TGA	-	*URI1*	0	0	rs3840928	0	A,B
chr19	49657711	CAT	-	*HRC*	0	0	rs66501117	0	A,B
chr21	47707039	-	AAAAAA	*YBE*	0	0	rs71318058	0	B

Table 4.10. Variants shared between all affecteds, that were identified using the BWA-Dindel pipeline. All were predicted as disease causing by MutationTaster. Columns titled EVS, 1000G, Control list give the variant allele frequency as listed in the exome server project (5400), 1000 genomes, and the control list respectively. All positions based on HG19 reference.

4.5 Discussion

I identified potentially damaging variants, segregating with disease in all three families. In particular, I was able to identify potentially disease causing variants in the HSP and DCM pedigrees. Variants in all three pedigrees were identified using whole exome enrichment and NGS, in combination with widely used sequence analysis and variant filtering methods. I was unable to identify any potentially disease causing variants in plausible candidate genes in the AVSD pedigree. There are various possible reasons for this (Discussed below). A detailed discussion of the results obtained for the DCM and HSP families will be given below, followed by a discussion on reasons why I think potentially disease causing variants may not have been recognised in the other two pedigrees.

DCM variants

The *ANKRD20A1* is a little known gene which is part of the ankyrin repeat domain 20 family (http://www.genecards.org). However, it is related to the *ANKRD1* gene which has been shown as a candidate gene for DCM (Moulik *et al.*, 2009). For example, the study by Moulik *et al.* (2009) screened 208 DCM patients for variants in the *ANKRD1* gene. The study identified three missense mutations. Functional studies indicated that these variants result in differential stretch-induced gene expression.

I identified a splice site variant in the *ANKRD20A1* gene (Figure 4.10), which due to previous reports of the influence of *ANKRD1* genes on DCM, may be of importance.

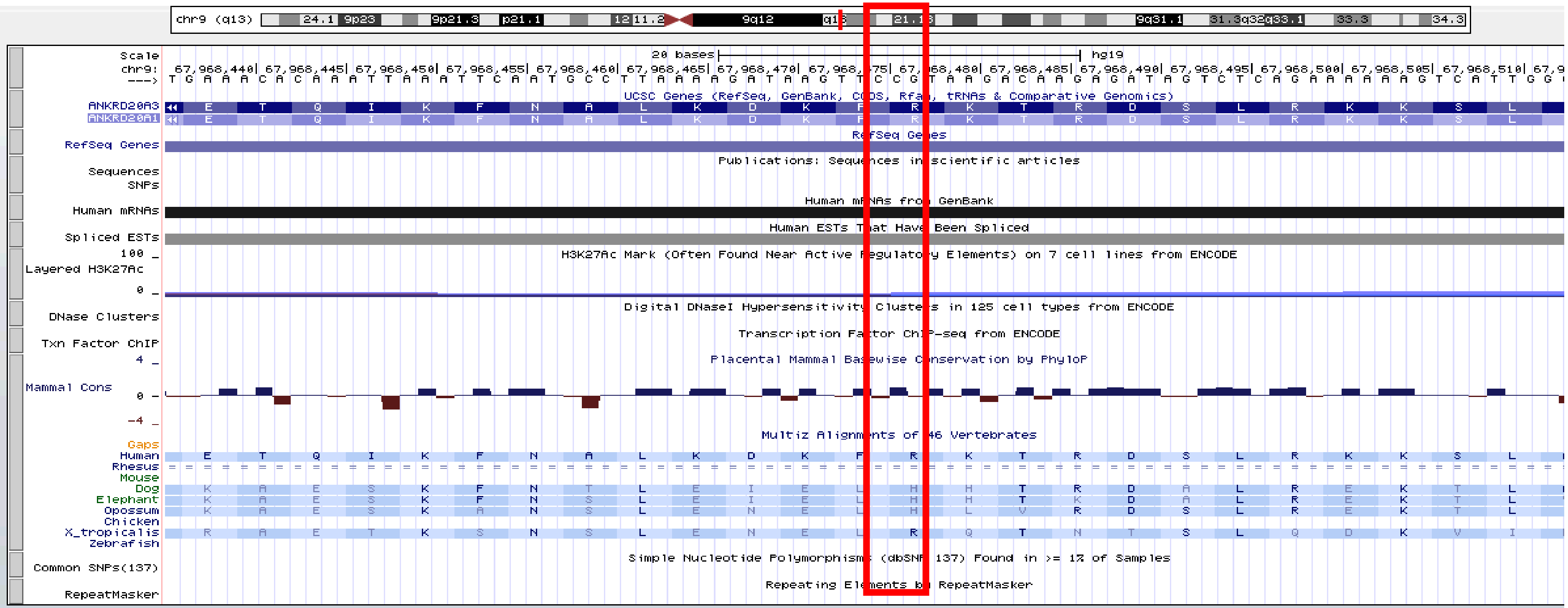

chr9 (q13)
24.1
9p23
9p21.3
p21.1
12
11.2
9q12
q1
21.1
9q31.1
31.3
q32
q33.1
33.3
34.3
Scale
chr9:
20 bases
hg19
67,968,440
67,968,445
67,968,450
67,968,455
67,968,460
67,968,465
67,968,470
67,968,475
67,968,480
67,968,485
67,968,490
67,968,495
67,968,500
67,968,505
67,968,510
T G A A A C A C A A A T T A A A T T C A A T G C C T T A A A A G A T A A G T T C C G T A A G A C A A G A G A T A G T C T C A G A A A A A A G T C A T T G G
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ANKRD20A3
ANKRD20A1
E T Q I K F N A L K D K F R K T R D S L R K K S L
RefSeq Genes
Publications: Sequences in scientific articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
4
Mammal Cons
0
-4
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 4.10. Splice site variant identified in the *ANKRD20A1* gene in the DCM pedigree.

HSP variants

I identified five variants of potential interest in members of the HSP family. All the variants represent single base substitutions and are located in the affected siblings, but not in the unaffected mother. Although not predicted as potentially disease causing by MutationTaster, the *FAM111B*, *AGAP6,* and *CNTNAP3B* genes are of interest as variants were identified in these genes in both my cases and in the cases of my collaborators. Due to the rarity of HSP, identifying shared genes containing non-synonymous variants in both unrelated pedigrees greatly increases their candidacy as possible candidates for disease. These three genes will be discussed in more detail below.

Firstly, I identified a non-synonymous variant in the *FAM111B* gene. This missense variant results in a p.Tyr591Asp change and is conserved across different species (Figure 4.11). The missense variant identified by my collaborators results in a p.Arg597Gly change. In both families the variant was present in only the affected individuals and had not been previously listed in public SNP databases. The *FAM111B* gene is also known as a Cancer-associated Nucleoprotein and belongs to the *FAM111* family (http://genome.cse.ucsc.edu; 23 December 2012). Very little is known of the gene, but it is likely to be an enzyme with peptidase cysteine/serine trypsin-like functions (Dr. Sébastien Kury; *perscomm*). The functional consequences of aberrations in this gene have not been previously identified, and Dr. Elise Glen (*perscomm;* University of Newcastle) has shown that it is expressed in the skin, and not in the liver.

Secondly I identified a non-synonymous single base substitution in the *AGAP6* gene. This missense variant is in a conserved region of the gene and results in a p.S260I change (Figure 4.12). As with *FAM111B*, a non-synonymous variant was also observed in the *AGAP6* gene in the HSP case provided by my collaborators. This gene is officially known as "*ArfGAP* with GTPase domain" (http://www.ncbi.nlm.nih.gov/gene/414189) and is a putative GTPase activating protein (http://www.uniprot.org/uniprot/Q5SRD3). The gene is of unknown function and no human genetic condition is known to result from mutations in the gene.

Finally, I identified a non-synonymous, missense variant within the *CNTNAP3B* in both my cases and the case from my collaborators. The variant is in a conserved region and results in a p.A2S change (Figure 4.13). Unfortunately, as with the last two genes, very little appears to be known about the function of *CNTNAP3B* besides that it may be involved in cell adhesion processes (http://www.uniprot.org/uniprot/Q96NU0). As mentioned previously, although very little is known about the functions of these three genes, given that this is a very rare disease, and affected individuals in both families have non-synonymous variants in the same genes, indicate that these genes could be related to HSP in these two families.

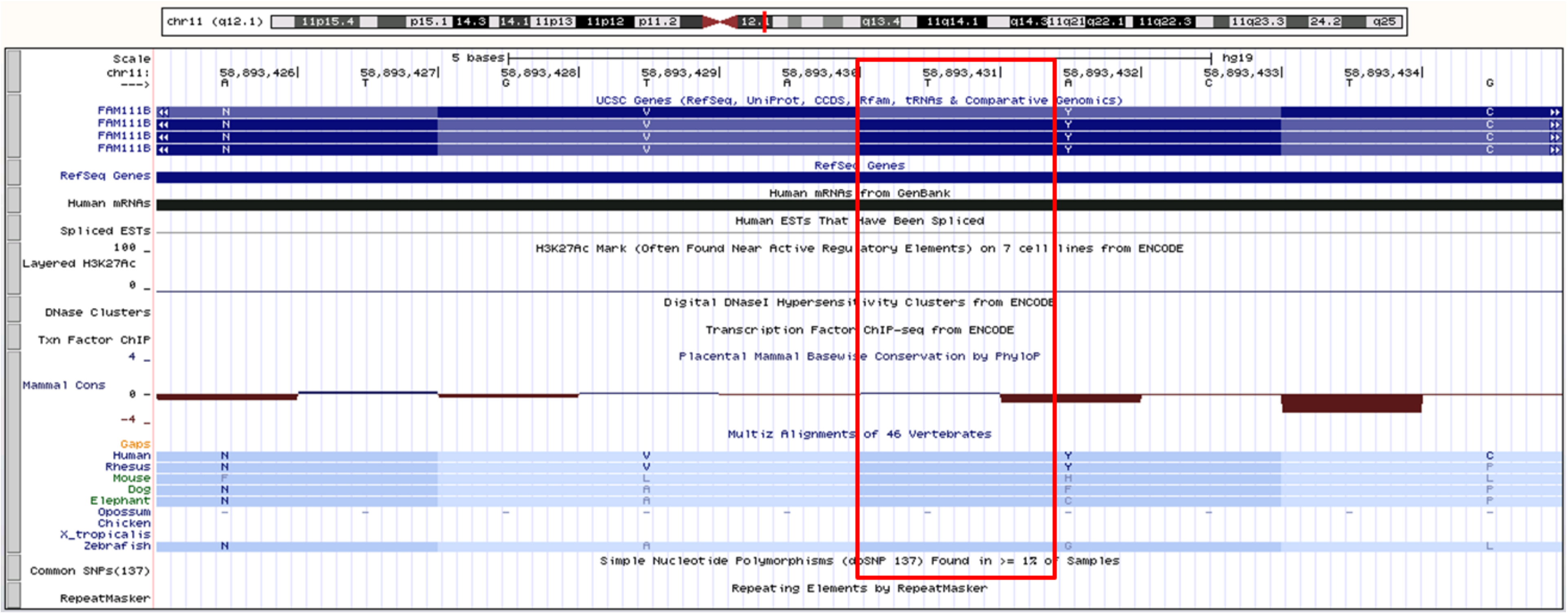


Figure 4.11. p.Tyr591Asp variant identified in the *FAM111B* gene in the HSP pedigree.

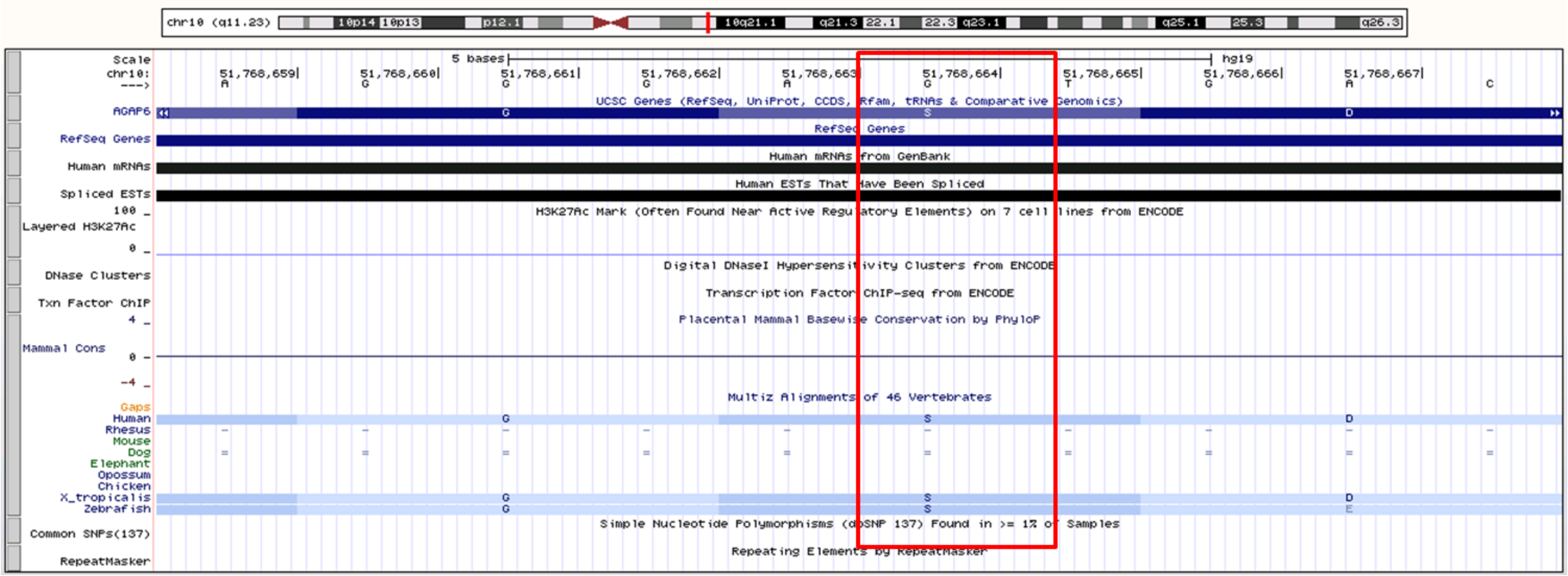

chr10 (q11.23)
10p14 10p13 p12.1 10q21.1 q21.3 22.1 22.3 q23.1 q25.1 25.3 q26.3
Scale
chr10:
--->
5 bases
hg19
51,768,659
51,768,660
51,768,661
51,768,662
51,768,663
51,768,664
51,768,665
51,768,666
51,768,667
A G G G A G T G A C
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
AGAP6
G S D
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
100 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0 _
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
4 _
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0 -
-4 _
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
G S D
G S D
G S E
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 4.12. p.S260I variant identified in the *AGAP6* gene in the HSP pedigree.

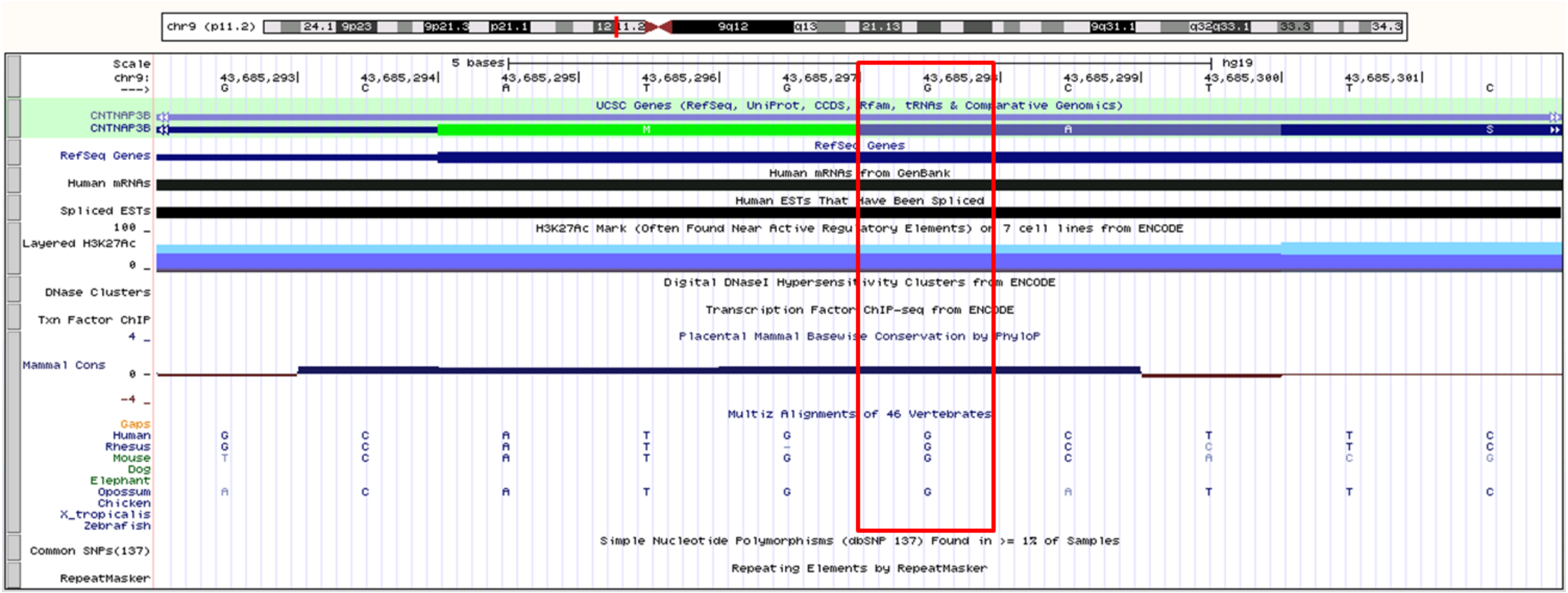

chr9 (p11.2)
24.1
9p23
9p21.3
p21.1
12
11.2
9q12
q13
21.13
9q31.1
q32
q33.1
33.3
34.3
Scale
chr9:
--->
5 bases
hg19
43,685,293
43,685,294
43,685,295
43,685,296
43,685,297
43,685,298
43,685,299
43,685,300
43,685,301
G
C
A
T
G
G
C
T
T
C
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CNTNAP3B
CNTNAP3B
M
A
S
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4 _
Mammal Cons
0 -
-4 _
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Figure 4.13. p.A2S variant identified in the *CNTNAP3B* gene in the HSP pedigree.

Aside from these three genes, I was also able to identify variants in two additional genes of possible interest in the two affected cases. However, these genes were not shared with the case of my collaborators but due to the possible functions of these two genes, they still warrant further discussion.

The first is a known (rs141865425) missense variant resulting in a c.G713A SNP change, with a population frequency <1% (1000 genomes database). It is located in a conserved region of the *BLK* gene. *BLK* belongs to the Src family kinases, which are thought to function in the cell proliferation and differentiation pathways (Islam *et al.*, 1995; Zwollo *et al.*, 1998). In particular, *BLK* is expressed in B lymphoid cell lines (Dymecki *et al.*, 1990), and in immature T cell lines (Islam *et al.*, 1995). More recently, the gene has also been found to be expressed in the spleen, liver, leukocytes, ovary, muscle and testis (Appel *et al.*, 2002).

Interestingly, the *BLK* gene is located in the 8p22-q23 chromosomal region thought to contain a gene responsible for Keratolytic Winter Erythema, KWE (Starfield *et al.*, 1997; Appel *et al.*, 2002). KWE is an autosomal dominant skin disorder resulting in erythema, keratosis and peeling of the palms and soles (Appel *et al.*, 2002). Many of the features of KWE are shared amongst the family suffering from HSP, in particular the keratosis and sclerosis of the palms of the hands and soles of the feet (Weary *et al.*, 1969; Lee *et al.*, 2012).

A study performed by Starfield *et al.* (1997) identified a region on chromosome 8 which was linked to KWE. The study involved a German family with 20 affected and 14 unaffected individuals. A panel of 230 genome wide, evenly spaced microsatellite markers was used to identify regions of linkage. Appel *et al.* (2002) designed and sequenced 7 BAC clones spanning the linkage region for KWE that was identified by Starfield *et al.* (1997). The BAC clones were used to identify a total of 12 transcripts covering the linkage region, one of which corresponded to the *BLK* gene. Direct sequencing of the gene was carried out using the individuals from the German pedigree in Starfield *et al.* (1997), and variants were subsequently identified. However, they were unable to identify any potentially pathogenic mutations in the KWE patients. The functional implications of the *BLK* gene, in particular its presence in the linkage region for KWE, make it a good candidate for HSP in this family. Although it is a SNP, it has a very low population frequency (<1%), so it may still be of interest with regard to HSP in this pedigree.

Finally I identified a missense c.C1889T change in the *ALOXE3* gene, which was predicted as disease causing by MutationTaster. *ALOXE3* is one of the five active LOX genes that are expressed predominantly in keratinised epithelia and functions in keratinocyte differentiation (Yu *et al.*, 2003). Variants within the *ALOXE3* gene have previously been reported to cause Non-bullous Ichthyosiform Erythroderma, NIE (Jobard *et al.*, 2002). Icthyoses comprise a heterogenous group of disorders characterised largely by scaly skin, with NIE in particular being characterised by hyperkeratosis and displaying an autosomal recessive pattern of inheritance (Oji and Traupe, 2006). By analysing 8 NIE patients from 6 families, Jobard *et al*. (2002) identified 3 nonsense mutations and a frameshift deletion shared by all the patients, and not found in 120 control individuals.

As with the *BLK* variant, the variant I identified in *ALOXE3* is a SNP, but it does have a very low population frequency. The possibility of mutations within the *ALOXE3* gene causing NIE, and the overlapping features of NIE with HSP, I think make this a potentially interesting variant.

Reasons for not identifying potentially disease causing variants

As mentioned previously, there are various reasons why I may not have identified any potentially disease causing variants in plausible candidate genes in the pedigree where cases presented with AVSD. Many of these reasons involve challenges of using exome capture in the detection of causative variants, and include, for example, sample choice, various technical limitations of target capture and sequencing, and in obtaining sufficient coverage. All these issues will be discussed in detail below.

Sample selection

Deciding which individuals, and of course how many, are to be sequenced in a pedigree is a very important consideration in exome sequence studies. When searching for very rare alleles, it may not be necessary to sequence all the affected individuals within a pedigree. In these cases Bamshad *et al.* (2011) suggest that because of the high probability of identity-by-descent, sequencing only two distantly related individuals within a pedigree could provide enough information to identify the disease causing variants. As an example, the exomes of two siblings were sequenced in an attempt to identify the causative variants for an inherited lipid metabolism disorder called hypobetalipoproteinemia (Musunuru *et al.*, 2010). As the disease is inherited in an autosomal recessive fashion, the investigators restricted their search to homozygous, novel variants present in both individuals and not present in the dbSNP database. Ng *et al.* (2010) also used NGS methods to discover the gene responsible for Miller syndrome by sequencing the exomes of only four unrelated individuals. The study identified two variants within the *ANGPTL3* gene, which were either not present in, or heterozygous in, 38 control exomes. However, both of these examples represent studies on recessive disorders which have proved more successful (Bamshad *et al.*, 2011).

Nevertheless, NGS studies on dominant disorders have proved successful, in particular when searching for causative *de novo* variants in dominant Mendelian disorders by sequencing of parent-offspring trios (Bamshad *et al.*, 2011). For example, this study design was successfully implemented to identify the causative variants in ten patients suffering from unexplained mental retardation (Vissers *et al.*, 2010). The study used the sequenced exomes of parent-offspring trios and identified an average of 21 755 variants per individual. Variants were further prioritised by selecting only non-synonymous and splice site variants and removing all those present within dbSNP and an in-house variant database. Finally, all remaining inherited variants were removed, resulting in a final list of 51 variants. Thirteen of the remaining variants could be validated via Sanger sequencing, 9 of which were present in 7 of the affected individuals, and absent in 1 664 controls. All 9 variants occurred in different genes, four of which displayed evidence for having a causal link to mental retardation in model organisms and protein interaction studies.

Data analysis issues of exome sequencing

Following sequencing, a base calling algorithm is used to determine the nucleotides from the intensity files produced by the sequencer (Nielsen *et al.*, 2011). Some of the main difficulties involved in base calling, for which all base callers have to correct, are phasing, pre-phasing and decreased signal intensity with each cycle (Ledergerber and Dessimoz, 2011). Phasing occurs when a sequence fails to add a base during a cycle, while pre-phasing is a term used to describe the situation where multiple bases are added during one cycle. An additional issue is that of cross-talk which refers to the overlap in emission spectra of the four fluorescent labels, which can impede identification of the correct base (Coonrod *et al.*, 2012). Reducing the error rate of base calls is important as it effects the downstream analyses and may result in a sequence not being aligned to the reference, being aligned to the incorrect position along the reference, or could result in false variant calls (Malhis *et al.*, 2009; Nielsen *et al.*, 2011).

Many programmes are available to align reads to a reference, such as Bowtie-2 (Langmead and Salzberg, 2012), NovoAlign (http://www.novocraft.com), BWA (Li and Durbin, 2009) and GSNAP (Wu and Nacu, 2010). Each of these programmes varies in its ability to correctly align reads to a reference (Li and Durbin, 2009; Wang *et al.*, 2011; Pattnaik *et al.*, 2012). For example Wang *et al.* (2011) compared the performance of various alignment algorithms by calculating the percentage of reads the program was able to map to the reference. The best performing program in their study was SHRiMP which aligned 81.23% of the reads to the reference, whereas the worst performing program was RMAP which only aligned 55.98% of the reads to the reference. Pattnaik *et al.* (2012) found that Bowtie was much faster than NovoAlign, but that it was only able to align 54.18% of the reads to the reference, whereas NovoAlign aligned 85.47% of the reads to the reference.

The alignment step has obvious implications for the accurate identification of variants, and it is important for these programmes to produce accurate read alignment quality scores as these can later be used by the variant caller (Nielsen *et al.*, 2011). However, accurate alignment does present with various difficulties, such as distinguishing true alignments from amongst multiple alignments (Wang *et al.*, 2011), distinguishing sequencing errors from real genomic differences (Nielsen *et al.*, 2011), and the fact that some areas of the genome are just difficult to align to, in particular those areas displaying a high level of inherent diversity within a population (Albers *et al.*, 2011; Nielsen *et al.*, 2011). For example, along homopolymer stretches where the indel polymorphism rate within a population is higher than in other genomic regions (Albers *et al.*, 2011). Regions of the genome containing high numbers of indels are difficult to align to (Harismendy *et al.*, 2009; Albers *et al.*, 2011; Coonrod *et al.*, 2012), where the presence of indel variants within the reads has been shown to increase both false positive and false negative calls (DePristo *et al.*, 2011).

Once reads have been aligned to the reference, variants can be identified as deviations in the reads from the reference sequence. The difficulty in this step involves accurately distinguishing the true genetic variations from the errors produced in sequencing or alignment (Shen *et al.*, 2010; Wang *et al.*, 2011). There are many programmes available to call genetic variants from aligned reads, such as Varscan-2 (Koboldt *et al.*, 2012), Dindel (Albers *et al.*, 2011), Samtools/bcftools (Li *et al.*, 2009) and GATK (McKenna *et al.*, 2010). As with the different alignment programmes, the use of different variant callers can have a marked effect on sensitivity (Ji, 2012; Pattnaik *et al.*, 2012). In conjunction with the bowtie aligner, using Samtools as a variant caller, Pattnaik *et al.* (2012) was able to match only ~40% of their identified variants to the dbSNP database, whereas by using GATK over 80% of the identified variants matched the dbSNP database.

Many studies have used similar methods to those which I used to identify the genetic variants responsible for different diseases (Ng *et al.*, 2009; Johnson *et al.*, 2010b; Krawitz *et al.*, 2010; Musunuru *et al.*, 2010; Ng *et al.*, 2010b; Wang *et al.*, 2010a; Norton *et al.*, 2011), including various in-house studies (Dickinson *et al.*, 2011; Horvath *et al.*, 2012; Pfeffer *et al.*, 2012; Pyle *et al.*, 2012). For example, in the study carried out by Pyle *et al.* (2012). However, there were some differences between the methods. The key difference between the methods employed by Pyle *et al.* (2012), and those employed here is in the use of different variant calling software. The study by Pyle *et al.* (2012) made use of the BWA aligner and Varscan variant caller to identify single nucleotide substitutions. Conversely, I made use of the NovoAlign aligner and Samtools/bcf tools variant caller to identify single nucleotide substitutions. However, in chapter 3 I demonstrated that, for the identification of single nucleotide substitutions, the NovoAlign-Samtools analysis pipeline is more sensitive than the BWA-Varscan analysis pipeline.

Technical issues of exome sequencing

The accuracy of variant calls is often seen to be strongly affected by the base quality score and sequencing depth (Nielsen *et al.*, 2011; Pattnaik *et al.*, 2012). In particular the depth of coverage has been shown to have a large effect on the false positive rate in Illumina sequence reads (Wang *et al.*, 2008; Harismendy *et al.*, 2009). A lack of sufficient coverage would also lead to increased numbers of false negatives and the disease causing variant not being identified (Zhi and Chen, 2012). Harismendy *et al.* (2009) demonstrated that at a sequence depth of ~10 fold, Illumina sequences have a false positive rate of 0.7, while at a coverage of ~68 fold this drops to only 0.1. This is also true for indel variants where sensitivity has been shown to rise from less than 0.85 to more than 0.95 at coverage depths of 10 and 20 respectively (Qi *et al.*, 2010). However, as >95% of the target bases were covered more than 10 fold, I think that the targets were sufficiently covered in the present study. See also chapter 3 where I assessed the affect of coverage depth on sensitivity.

A major flaw of exome sequencing is that not all of the coding regions are actually covered by commercial targeting kits (Asan *et al.*, 2011; Parla *et al.*, 2011; Sulonen *et al.*, 2011). The problem arises in trying to define a set of targets that would encompass the exome, as not all of the protein coding sequences making up the human genome are known (Bamshad *et al.*, 2011). Two widely used exome capture kits include the Agilent SureSelect kit and the NimbleGen kit.

Parla *et al.* (2011) assessed the ability of both the Nimblegen (26.2Mb targets) and Agilent (37.6Mb targets) kits to capture known coding regions based on their intended targets of the CCDS. They found that the Agilent kit covered 97% of the CCDS targets, whereas the Nimblegen kit only covered 88% of the CCDS targets. However, Asan *et al.* (2011) found that a higher proportion of reads could be mapped to the reference sequence using the Nimblegen technology (>10% higher), rather than the Agilent technology. This superior target enrichment using the Nimblegen targeting kit has also been confirmed in other studies (Clark *et al.*, 2011; Sulonen *et al.*, 2011)

The performance of both the Nimblegen and Agilent kits in variant identification studies was tested by Asan *et al.* (2011). They found that at a sequencing depth of 30 fold, Nimblegen displayed a higher sensitivity, and they were able to identify 12 400 variants in the targeted coding regions common to both kits, whereas when using the Agilent kit only 12 000 SNPs were identified in these regions. However, they also found that the Agilent kit detected 13 500 coding SNPs outside of these common coding regions, whereas the Nimblegen kit only detected 12 600 coding SNPs outside of these regions. They attribute this to the higher capture efficiency of the Nimblegen kit, and the larger area (~4Mb larger) captured by the Agilent kit.

However, at 20 fold coverage Sulonen *et al.* (2011) found that both Nimblegen and Agilent kits could provide comparable, highly sensitive SNP calls (>97%) which they calculated by using the SNPs captured on the Illumina Human660W-Quad v1 SNP chip. This was further corroborated by Clark *et al.* (2011), who found a concordance of >99%

when SNP calls from both were compared to the Illumina 1M-Duo SNP chip. In chapter 3, I used the NovoAlign aligner and Samtools variant caller to identify variants within 19 exomes. On average, I was able to identify in excess of 98% of the on target markers on the Illumina 660W SNP chip.

Issues of filtering in exome sequencing

Exome sequencing studies often identify many thousands of variants. For example, I identified >88000 variants in the HSP pedigree and >222000 variants in the DCM pedigree. Due to these very large numbers of variants, various filtering steps are often used to reduce this number to a more manageable size. Often, one of the first steps used to reduce down the number of identified variants is the removal of all the variants occurring outside of the target regions (Dickinson *et al.*, 2011; Horvath *et al.*, 2012; Pfeffer *et al.*, 2012; Pyle *et al.*, 2012). The obvious danger with this technique is that the causative variant may lie outside of the target regions.

By selecting only variants present within the targeted regions, I was assuming that the causative variant is exonic. However, Cooper *et al.* (2010) estimated that up to ~14% of the mutations within the Human Gene Mutation Database (A database containing known genes responsible for human inherited diseases) are located within the intronic and regulatory regions of genes (Cooper *et al.*, 2010). There are also many examples where intronic variants and those found in 3' and 5'-untranslated regions have been found to affect disease (Scheper *et al.*, 2007; Chen *et al.*, 2010). For example, in the case of a Retinitis Pigmentosa (RP) where linkage mapping suggested the involvement of the *PRPF31* gene, extensive screening of the genes exons failed to identify the causative variant (Rio Frio *et al.*, 2009), and sequencing the entire *PRPF31* gene allowed Rio Frio *et al.* (2009) to identify a deep intronic single base substitution causing RP in this family. Also, Scheper *et al.* (2007) provides examples of various inherited diseases caused by mutations within the 5` UTR's of genes, such as in hereditary hyperferritinaemia.

As well as selecting only on-target variants, I also removed variants recorded in public databases such as 1000 genomes. For a single European sample it is expected that between 74% - 95% of all identified variants will be present in a public database (Bentley *et al.*, 2008; DePristo *et al.*, 2011; Coonrod *et al.*, 2012). This value has obviously increased over time, and now in excess of ~95% of variants should be expected to be present in public databases. More than 96% of the variants I identified in all three families were already recorded in public databases. Of course, this method assumes that the variants present in these databases are common in the population and cannot therefore be the causative variant for a rare disorder, and often results in a considerable reduction in the numbers of variants from the candidate list (Ng *et al.*, 2009; Wang *et al.*, 2010a; Bamshad *et al.*, 2011; Norton *et al.*, 2011; Coonrod *et al.*, 2012).

However, there is a risk that the causative variant may be present in the population and therefore also in these databases, albeit at a low frequency. This is particularly relevant in the case of recessive disorders where carriers do not present with the disease phenotype (Bamshad *et al.*, 2011). Therefore, it is becoming more popular to employ minor allele frequency thresholds when removing variants that are present in these databases (Bamshad *et al.*, 2011; Stitziel *et al.*, 2011).

Applying a base coverage threshold to remove poorly supported variant calls is also often used as a filtering criterion. Coverage values falling outside of the normal range, i.e. excessively high or low coverage at a particular position, may result in false positive or negative calls (Bentley *et al.*, 2008; Coonrod *et al.*, 2012). As well as total coverage at a position, a minimum coverage of the variant allele has also been shown to be valuable in identifying false positive and negative calls (Mokry *et al.*, 2010).

A further filtering approach which can be implemented is to incorporate a prediction programme to estimate the potential impact of variants on protein function (Jordan *et al.*, 2010; Bamshad *et al.*, 2011). In the current study, the MutationTaster prediction programme was used to assess the pathogenicity of variants. MutationTaster is a free programme which uses evolutionary conservation, annotation and structural information to assess the impact of a particular variant (Schwarz *et al.*, 2010). Many other programmes also are available to predict potential variant pathogenicity, such as SIFT (Kumar *et al.*, 2009) and PolyPhen-2 (Adzhubei *et al.*, 2010).

One of the major concerns regarding these prediction algorithms is the variations in sensitivity and specificity achieved between them (Chan *et al.*, 2007; Hicks *et al.*, 2011). Hicks *et al.* (2011) assessed the sensitivity and specificity of various prediction programmes, and highlighted a large variation in the results obtained using the different algorithms. For example, Polyphen2 and SIFT achieved sensitivity values of 0.9 and

0.85 respectively and specificity values of 0.40 and 0.52 respectively. In another study the accuracy of predictions between four different algorithms was assessed. The worst performing method was a program called A-GVGD (Tavtigian *et al.*, 2006), which achieved a sensitivity of 72.9, whereas Polyphen was able to achieve a sensitivity of 83.3 (Chan *et al.*, 2007). Although prediction programmes may not correctly predict variant effects, they are still considered a useful means of prioritising variants in sequencing studies (Karchin, 2009; Jordan *et al.*, 2010).

Another common filtering approach is to focus only on those changes which alter the amino acid. By using the Human Gene Mutation Database, Kryokov *et al.* (2007) calculated that up to 20% of all missense mutations could result in a complete loss of protein function. They also estimated that up to 53% of all *de novo* missense mutations can be considered as mildly deleterious, which they defined as mutations which do effect, but not completely eliminate, protein function. Although removing all of the synonymous and common variants may provide an effective means of reducing the overall number of variants, there is a chance that these may include the disease causing variants (Ku *et al.*, 2011).

4.6 Conclusions/Future work

In this chapter I identified variants potentially causing HSP and DCM. In the HSP pedigree, these include variants in the *BLK*, *ALOXE3*, *FAM111B*, *AGAP6*, and *CNTNAP3B* genes. Of particular interest are the variants in the *FAM111B, AGAP6*, and *CNTNAP3B* genes. Although, current literature and knowledge regarding the effect of variants in these genes suggest that they could potentially influence HSP in this family, functional analyses will be required to determine their role in disease. I also identified a variant in the *ANKRD20A1* gene which could be responsible for disease in the DCM pedigree. There are various reasons why I was not able to identify any potentially disease causing variants in plausible candidate genes in the AVSD pedigree.

In particular, the first three stages of variant filtering (i.e. the removal of off-target variants, selecting only those variants shared amongst affected individuals, and removing the homozygous changes) removed more than 90% of the identified variants. This suggests that these three steps are a very effective means of reducing variants down to a more manageable number, and should presumably make up the first three stages in any filtering approach. However, care should be taken when removing variants occurring outside of the targets, as they may still be important.

Therefore, any future work should alter the filtering criteria I used, or possibly more importantly, use a whole genome sequencing approach to capture more of the genome. However, this approach is currently limited by current cost and resource requirements. Furthermore, it may prove valuable to search for other kinds of variants, such as copy number variants. I only identified single nucleotide and indel variants, however, it is possible that disease in the DCM and AVSD families are caused by genomic variants such as CNV's. See chapter 1 and 6 for a detailed discussion on these issues.

Chapter 5. Identifying disease causing indels using targeted next generation sequence data from patients with congenital cardiovascular disorders

5.1 Aim

This study was conducted as part of an international collaborative project which aimed to identify rare variants potentially causing certain congenital heart malformations, in particular those characterised by ventricular hypoplasia. This was done by sequencing selected genes in a group of patients with various congenital malformations. My role within the project was to analyse targeted NGS data to identify potentially disease causing insertion/deletion (indel) events in these patients.

5.2 Introduction

5.2.1 Sample Origin

One hundred and thirty three patient samples were provided by six centres located in The Netherlands (Academic Medical Center, Amsterdam and Leids Universitair Medisch Centrum, leiden), England (The University of Newcastle, Newcastle Upon Tyne), Belgium (Katholike Universiteit, Leuven, University of Leuven, Leuven), and Germany (Max Planck Institute for Molecular Genetics and the Max Delbrück Center for Molecular Medicine in Berlin). These patients suffer from a range of congenital cardiac disorders (Details on sample phenotyopes are provided in table 5.1) that are characterised by underdevelopment (hypoplasia) of either the left or the right ventricular chamber.

Right Ventricular hypoplasia/malformation:

Double inlet left ventricle - 21

Tricuspid atresia - 25

Right ventricular hypoplasia - 6

Pulmonary atresia with intact ventricular septum - 12

Ebstein's anomaly - 25

Left Ventricular hypoplasia/malformation:

Hypoplastic left heart syndrome - 8

Mitral valve atresia - 4

Left ventricular hypoplasia - 6

Other:

Noncompaction - 19

Univentricular heart - 7

Table 5.1. Categories, and subcategories, of congenital cardiac malformation that the 133 patients used in this study suffered from. The numbers represent the number of patients.

5.2.2 Phenotypes

All cases suffer from what are broadly termed "univentricular heart" defects, that comprise a range of malformations which are not easy to classify (Khairy *et al.*, 2007). All these malformations have a poor prognosis and if left untreated, survival into late adulthood is rare (Hager *et al.*, 2002). The cases in this study were categorised as suffering from right ventricular malformations and left ventricular malformations. In addition there is a group of 26 cases classified as "Other", because they did not fit strictly into any of the subcategories.

Left ventricular hypoplasia comprises a range of congenital heart abnormalities characterised by the severe underdevelopment of the left side of the heart and which often prove to be lethal (Trivedi *et al.,* 2011; Hickey *et al.,* 2012). It is frequently associated with obstruction to left ventricular outflow, where the degree of hypoplasia is proportional to the degree of obstruction (Hickey *et al.,* 2012). In severe cases, the left ventricle is unable to support systemic circulation, and the only options available for long term survival include neonatal heart transplant or a sequence of complex open-heart operations during infancy (Fruitman, 2000; Trivedi *et al.*, 2011; Hickey *et al.*, 2012). It is generally accepted that lack of flow during embryonic development plays a critical role in the pathogenesis of left ventricular hypoplasia; in models where left sided flow can be readily modelled (eg the chick embryo), restriction of left sided flow reproducibly results in hypoplasia of left heart structures. Therefore, genes particularly involved in the development of critical left heart structures such as the mitral valve and aortic valve might be considered particularly good candidates for involvement in left heart hypoplasia. In addition, however, some studies have shown the presence of mutations in transcription factors critical to the specification of left ventricular myocardium in patients with left heart hypoplasia (Grossfeld, 2007; Hickey *et al.*, 2012). Numbers, however, remain small due to the rarity and serious nature of this group of phenotypes, which remain the CVM phenotypes most likely to result in childhood death. Also, since the widespread availability of fetal cardiology services, the incidence of hypoplastic left heart syndrome, which is generally detectable using fetal echocardiography, has decreased due to termination of affected foetuses. As seen in table 5.1, even the pooled resources of a number of international congenital heart disease units resulted in the availability of relatively small numbers of patients with left ventricular hypoplasia.

Right ventricular (RV) hypoplasia describes a group of cyanotic congenital heart disease conditions characterised by a small right ventricle, and which can lead to congestive heart failure and cyanosis during infancy (Goh *et al.*, 1998). RV hypoplasia can be caused by the underdevelopment of one or more of a variety of structures on the right side of the heart, including the tricuspid valve, right ventricle (as a primary event), pulmonary valve, and the pulmonary artery (Van der Hauwaert and Michaelsson, 1971; Dib *et al.*, 2012). The degree of underdevelopment is highly variable, with very severe forms presenting in early infancy, while in the less severe forms the patient can survive to adulthood (Dib *et al.*, 2012). As is seen from Table 5.1, patients with RV hypoplasia from various causes were more readily collected from the collaborating centres. It is however important to be aware that the right-sided conditions studied in Table 5.1 are far from common; for example pulmonary atresia with intact ventricular septum represents only 1-3% of all congenital heart disease, and Ebstein's anomaly occurs in less than 1:20000 live births. Both right and left-sided phenotypes were therefore selected for rarity and severity; it was reasoned that selection of this group of patients would maximise the chances of finding rare variants of large phenotypic effect through NGS.

5.2.3 Indels and disease

Here I will define indels as deviations from the reference sequence where bases are either removed (deletions) or have been added (insertions). It is estimated that every individual harbours between 0.3 – 0.6 million indels, making these the second most common form of genetic variation, following SNPs (Levy *et al.*, 2007; Bansal and Libiger, 2011; Lemos *et al.*, 2012). Indels display very large variations in both distribution and size across the human genome, with sizes ranging from 1 to several 1000 base pairs (bp) in length (Bhangale *et al.*, 2005; Levy *et al.*, 2007; Wheeler *et al.*, 2008; Lemos *et al.*, 2012). For example, using 454 FLX sequence reads and a combination of the BLAT and cross_match (http://www.phrap.org/phredphrapconsed.html) programmes, Wheeler *et al.* (2008) were able to identify 222718 indels in a single individual, which ranged in size from 2 – 38896bp long (Table 5.2).

There is also a large amount of variation in indel frequency across different gene regions (MacArthur *et al.*, 2012). Bhangale *et al.* (2005) identified 2393 indels in a set of 330 targeted genes, and found that indels occurred more frequently in the 3'-UTRs than in the 5'-UTRs of genes. This pattern is explained by the greater tolerance to truncation close to the end of the coding regions (MacArthur *et al.*, 2012). However, despite this variation across different gene regions they identified very few indels in the coding regions of the genes, which they attributed to a strong negative selection on coding indels. The scarcity of coding indels in human genes is further corroborated by Wheeler *et al.* (2008), in which less than 1% of the 222718 indels they identified were located in coding regions.

Due largely to the difficulties involved in identifying indels using NGS methods (Bansal and Libiger, 2011), which are discussed in more detail below, far less is known about the effect of indels on genes than is known about SNPs (Cartwright, 2009; Mills *et al.*, 2011; Hu and Ng, 2012; Lemos *et al.*, 2012). This, despite indels of less than 20 base pairs long accounting for nearly one quarter of known Mendelian disease mutations (Hu and Ng, 2012). Therefore, a detailed knowledge of indel variation and distribution in patient samples would be very useful to understand their potential influence on disease (Mills *et al.*, 2011).

Length	Deletions	Insertions
1	664 (41.76)	397 (57.87)
2	309 (19.43)	68 (9.91)
3	188 (11.82)	39 (5.69)
4	185 (11.64)	79 (11.52)
5	69 (4.34)	26 (3.79)
6	29 (1.82)	17 (2.48)
7	14 (0.88)	11 (1.6)
8	20 (1.26)	6 (0.87)
9	12 (0.75)	4 (0.58)
10	9 (0.57)	7 (1.02)
11	12 (0.75)	0 (0)
12	17 (1.07)	0 (0)
13	8 (0.5)	3 (0.44)
>=14	54 (3.4)	29 (4.23)

Table 5.2. Lengths of insertions and deletions identified in 330 targeted genes. Values represent insertion and deletion counts, while the values in brackets represents the proportion. From Bhangale *et al.* (2004).

5.2.4 Indel identification using NGS data

The development of tools to accurately identify indels is a very important step in the search for the genetic causes of disease (Bansal and Libiger, 2011). Various different programmes have been developed to try and identify indels from NGS reads, including MAQ (Li *et al.*, 2008a), dindel (Albers *et al.*, 2011), and GATK (McKenna *et al.*, 2010). However, the performance and accuracy of these different programmes varies (Vallania *et al.*, 2010; Albers *et al.*, 2011). Using simulated data, Albers *et al.* (2011) compared the false discovery rate of indel calls using Dindel, Varscan and SAMtools. Dindel achieved the lowest false discovery rate of 1.56%, while Varscan had the highest rate of false discoveries, 16.67%. In my analysis in chapter 3, I calculated the sensitivity values for different indel calling pipelines and found large differences between them. Across 12 samples, the BWA-Dindel pipeline performed the best, achieving an average sensitivity of ~35%. Conversely, the NovoAlign-Samtools pipeline performed the worst, achieving an average sensitivity of <5%. This is considerably lower than the values achieved when identifying single nucleotide substitutions, highlighting the difficulties involved in identifying indels using NGS methods.

Despite this, many studies have been able to identify disease causing indels using these approaches (Wei *et al.*, 2011; Carmignac *et al.*, 2012; Drielsma *et al.*, 2012; Fuchs-Telem *et al.*, 2012; Pyle *et al.*, 2012; Wang *et al.*, 2012; Weterman *et al.*, 2012). For example Wei *et al.* (2011) identified an exonic, frameshift deletion in the DMD gene causing Duchenne muscular dystrophy. Target capture was performed using a Nimblegen custom capture array, and sequencing performed using an Illumina HiSeq2000. Sequence reads were aligned to the human genome using BWA and indels were identified using GATK. They identified a large deletion of exon 1 in the DMD gene that was present in all affected samples, but not found in 100 controls. In a further study by Pyle *et al.* (2012), the BWA-Dindel analysis pipeline was used to identify one and two base pair deletions in the *SACS* gene causing prominent sensorimotor neuropathy. The mutations were identified in two affected siblings and not present in 346 control samples, or in the 1000 genomes project.

Despite various successes, identifying indels using short read sequence data does present with various problems (Albers *et al.*, 2011). In particular, correctly mapping reads containing indels to the reference (Lunter and Goodson, 2011), particularly in cases where reads contain large insertions, is difficult (Albers *et al.*, 2011). Still further problems may include, an increased rate of indel false positive calls in highly polymorphic gene regions, and the presence of technological artefacts such as polymerase slippage during PCR amplification (Albers *et al.*, 2011; Bansal and Libiger, 2011).

5.2.5 Indel prioritisation

As with SNPs, programmes designed to locate indels will identify many thousands of variants per patient (Wei *et al.*, 2011). Therefore, there is a need for methods which could be used to filter these indel variants down to a more manageable number (Zia and Moses, 2011; Hu and Ng, 2012). As well as reducing the total number of potential indels that need to be validated, applying variant filters could help improve the specificity of indel calls (Albers *et al.*, 2011). As an example, Albers *et al.* (2011) suggest that, at the very least, indels should be required to be present on both the forward and reverse strands. Additionally, Mardis *et al.* (2009) suggest applying a coverage threshold to indel calls by, for example, only accepting those supported by at least 2 reads. Wei *et al.* (2011) prioritise indels by selecting only those which alter the protein and removing those present within the dbSNP, 1000 genomes and HapMap databases, as well as those present in an in-house list of controls. This is particularly relevant in the context of my study where I was expecting the variants to be rare and therefore not present, or present at a low frequency, in public databases.

One of the simplest means to prioritise indels may include selecting coding, frameshift indels (Hu and Ng, 2012), as these are reported to occur very rarely (Wheeler *et al.*, 2008); see also the 1000 genomes (http://www.1000genomes.org/) and Exome Variant Server (http://evs.gs.washington.edu/EVS/) databases. However, not all indels occurring in coding regions lead to a loss of function, some are functionally neutral, and indels occurring outside of the coding regions could also have a considerable impact on genes by, for example, altering splicing (Pagani and Baralle, 2004; Zia and Moses, 2011; Hu and Ng, 2012). Alternatively, Wei *et al.* (2011) suggest employing a disease database, such as the HGMD (http://www.hgmd.cf.ac.uk/), as a means of selecting indels which may already be known to cause disease. Another more complex means to prioritise indels is by using a prediction programme (Lemos *et al.*, 2012). However, as mentioned in chapter 4, Lemos *et al.* (2012) warn of discrepancies between the results produced by some of these programmes.

5.2.6 Indel validations

Even after prioritisation, validation rates for indels are lower than for single nucleotide substitutions (Weber *et al.*, 2002; Mardis *et al.*, 2009). For example, using PCR methods Weber *et al.* (2002) could only achieve a validation rate of 58% for indels of at least

2bp's long. For single base indels this rate dropped down to ~14%, and as single base indels are the most common type of indel, their low validation rate is an important issue. In another study, Mardis *et al.* (2009) used the MAQ aligner and Samtools variant caller to identify possible disease causing indels in a patient suffering from Acute Myeloid Leukemia. They identified 142 indels, of which they were only able to validate 23 (~16%).

5.3 Materials and Methods

5.3.1 Samples, gene selection, and sequencing

Centres in The University of Newcastle, Newcastle Upon Tyne (Prof. B Keavney), the AMC, Amsterdam (Dr Alex Postma), and the Max Delbrück Center, Berlin (Prof Sabine Klaasen), provided 133 samples. Centres in Newcastle and Amsterdam provided 67 samples each, while centres in Berlin provided 26 of the samples. Samples were obtained from individuals suffering from a range of congenital cardiovascular malformations.

Targeted genes were prioritised in four stages by the lead authors of the study in all three centres. Firstly, genes containing mutations which have previously been reported to cause human CVM were selected. Secondly, genes shown to be involved in CVM in mice or other model organisms were selected. Thirdly, genes known to participate directly in known regulatory gene networks for heart development were selected. Finally, genes involved in known gene networks for CVM, not necessarily directly, were selected. The final gene list consisted of 403 genes (Supplementary table 2.1). The number of genes represented on the array was limited by the size of the capture possible using the Agilent SureSelect system at the time the project began. Sequencing was performed in Amsterdam (LUMC, Leiden) using an Illumina Genome analyser IIx. More detail on the methods used here are provided in chapter 2.

5.3.2 Indel calling

The BWA aligner was used to map the reads to the human genome reference sequence (Build 37, hg19). Given the known lack of specificity of indel calls in NGS data, I decided to use two variant callers, Dindel (Albers *et al.*, 2011) and the Genome Analysis Toolkit (GATK) indel caller (McKenna *et al.*, 2010; DePristo *et al.*, 2011), focusing on the indels that were detected using both pipelines. It was hoped that this would increase the confidence of indel calls and decrease the amount of effort required in attempts at validation (using Sanger sequencing) of the many false positive variants that might be detected by one pipeline alone. Only reads that aligned uniquely to the reference and non-duplicate reads were selected using a combination of custom written Perl scripts and the Picard MarkDuplicates routine (http://picard.sourceforge.net/). Only the filtered indels called by both pipelines were validated using Sanger sequencing.

5.3.3 Indel filtering and annotations

As a first filtering step, all off-target variant calls were removed, and the filtered list submitted to wAnnovar (http://wannovar.usc.edu/) for indel annotations. Following annotations, only exonic, frameshifting and splice site indels were selected, and because I wanted to identify rare variants, all indels present in the EVS database, consisting of variant calls from 5400 human exomes, and 1000 genomes databases, comprising variant calls from multiple genomes, at a frequency exceeding 1% were removed (Figure 5.1). A control list of variants was also used as a filter. The control list contains indel variants from 114 unrelated exomes (Generated locally). All exome data comprising the control list had been sequenced in house using the Agilent 50Mb Whole Exome Targeting kit and the Illumina Genome Analyser IIx. The list was compiled by Dr. Helen Griffin (2012, *pers. comm*). All variants present within the control list at a frequency of more than 1% were removed. MutationTaster v20100416 (Schwarz *et al.*, 2010) was used to assess potential variant pathogenicity.

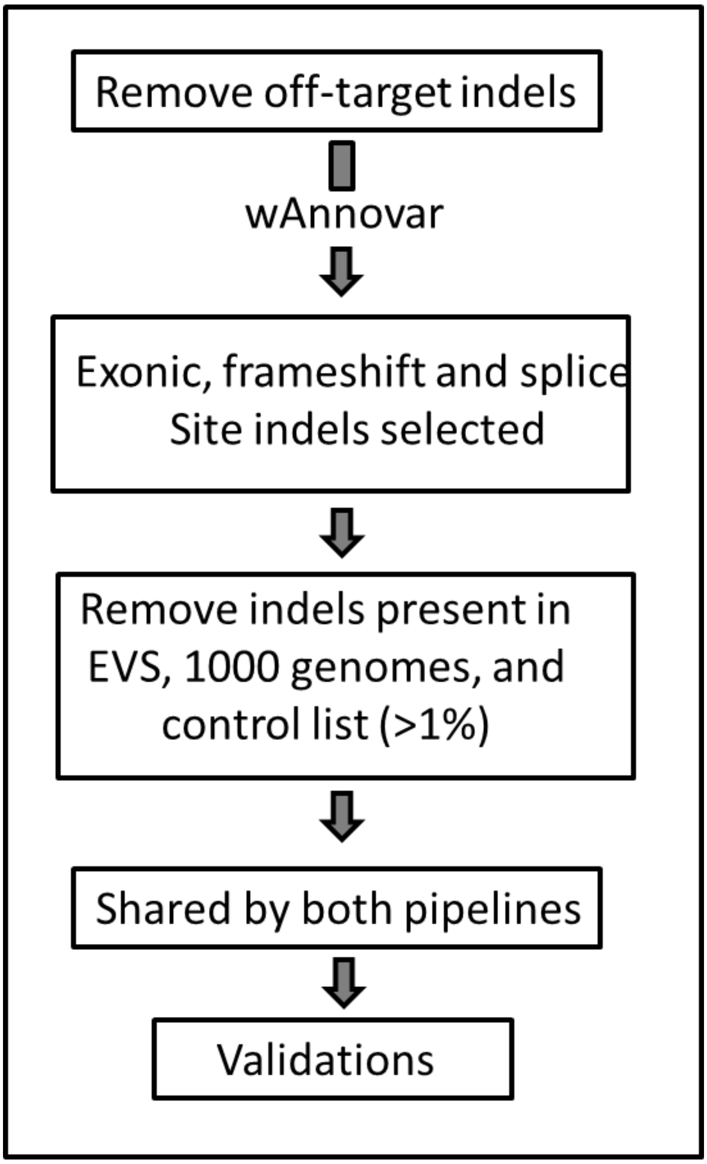

Remove off-target indels
wAnnovar
Exonic, frameshift and splice Site indels selected
Remove indels present in EVS, 1000 genomes, and control list (>1%)
Shared by both pipelines
Validations

Figure 5.1. Filtering steps used to prioritise indel calls.

5.4 Results

5.4.1 Alignment results

Average target base coverage ranged from 18 – 704x across all 133 samples. Between 85 - 97% of the target bases were covered at least once, and between 25 - 93% of the target bases were covered to a minimum of 10 fold (Appendix table 5.1).

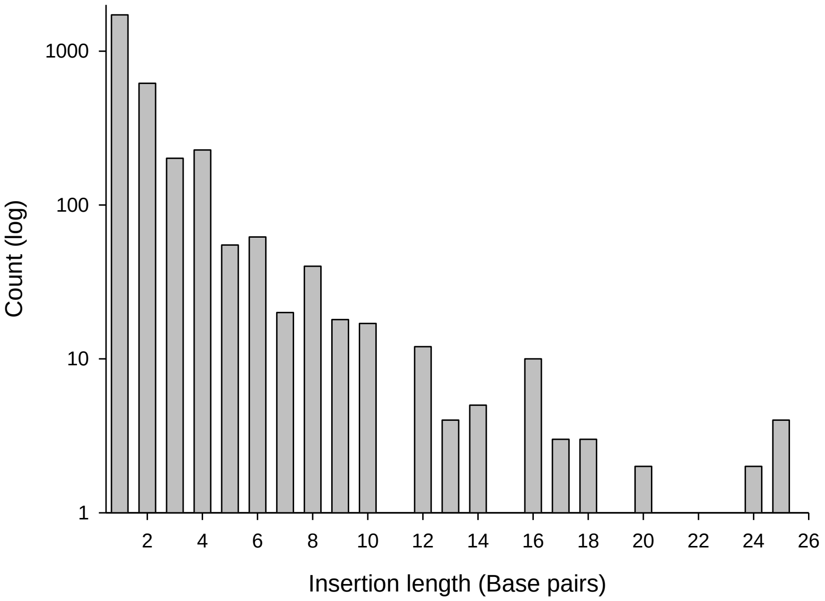
5.4.2 BWA-Dindel pipeline

Using the BWA-Dindel pipeline, an average of 696 non-unique indels were identified in each sample. This comprised 7223 unique indels, of which 3026 were insertions and

4197 were deletions. Insertions ranged in size from 1bp - 25bp’s long, with insertions of 1bp long being the most frequent (Figure 5.2A). Deletions ranged in size from 1bp - 40bp’s long, and as with the insertions, deletion lengths of 1bp were the most frequent (Figure 5.2B).

After filtering 317 indels remained (Appendix table 5.2).

5.2A



5.2B

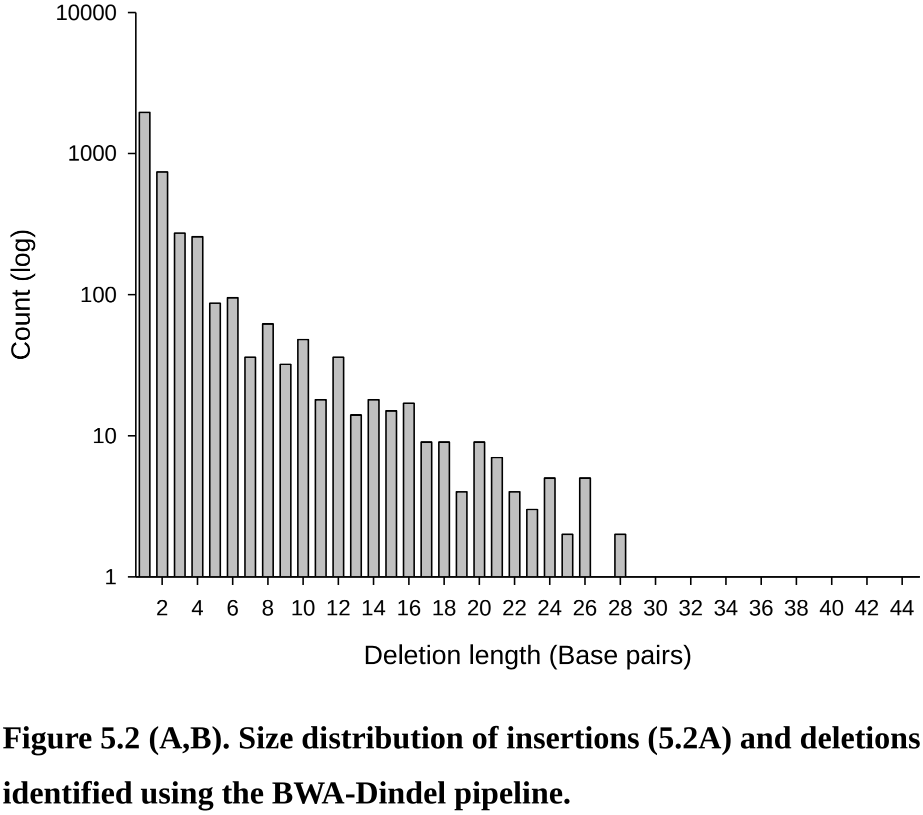


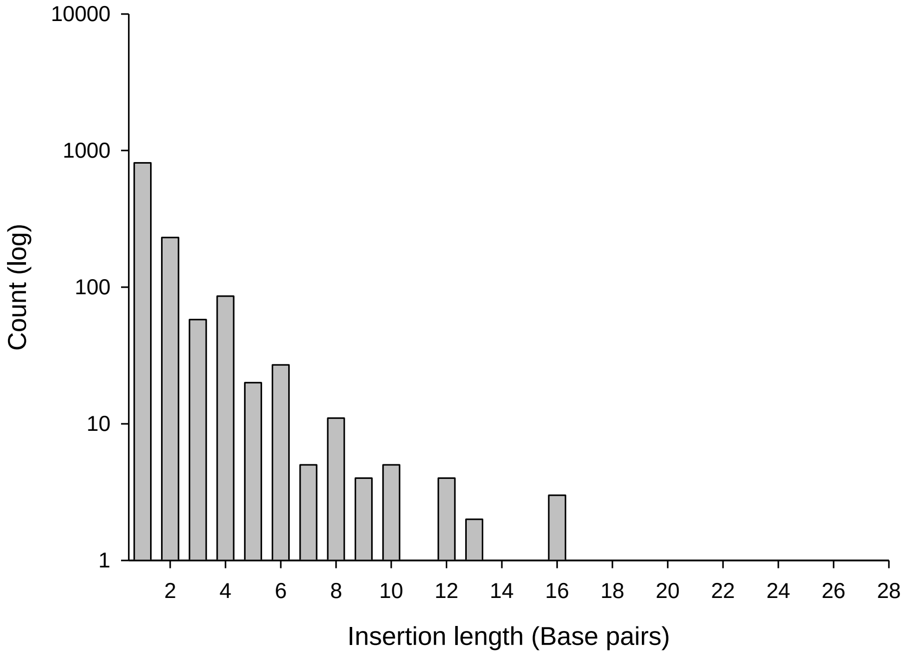
Figure 5.2 (A,B). Size distribution of insertions (5.2A) and deletions (5.2B) that I identified using the BWA-Dindel pipeline.

5.4.3 BWA-GATK pipeline

Using the BWA-GATK pipeline, an average of 526 non-unique indels were identified in each sample. This comprised 2873 unique indels, of which 1272 were insertions and 1 601 were deletions. Insertions ranged in size from 1bp - 24bp's in length, with 1bp insertions being the most frequent (Figure 5.3A). Deletions ranged in length from 1bp long - 40bp's long, with deletions of 1bp long being the most frequent (Figure 5.3B).

After filtering 35 indels remained (Appendix table 5.3).

5.3A



5.3B

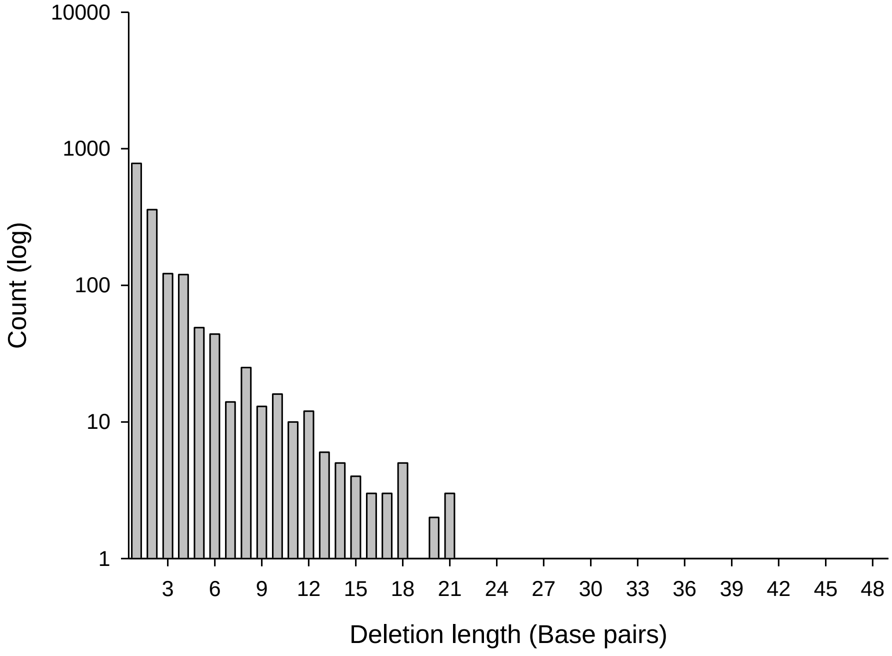


Figure 5.3 (A,B). Size distribution of insertions (A) and deletions (B) that I identified using the BWA-GATK pipeline.

5.4.4 Indels called by both pipelines

Of the 317 filtered indels from the BWA-Dindel pipeline and the 35 filtered indels from the BWA-GATK pipeline, 25 were identified by both (Table 5.3). Two hundred and ninety two were unique to the BWA-Dindel pipeline, and 10 were unique to the BWA-GATK pipeline (Table 5.4).

Chromosome	Position	Reference	Variant	Gene	Alteration
chr1	2235396	-	G	*SKI*	frameshift insertion
chr1	71418662	G	-	*PTGER3*	frameshift deletion
chr1	92185675	-	C	*TGFBR3*	frameshift insertion
chr1	120612003	GG	-	*NOTCH2*	frameshift deletion
chr1	201334355	T	-	*TNNT2*	frameshift deletion
chr1	202407190	T	-	*PPP1R12B*	frameshift deletion
chr2	66739381	-	T	*MEIS1*	frameshift insertion
chr2	121747069	-	A	*GLI2*	frameshift insertion
chr2	211179765	-	T	*MYL1*	frameshift insertion
chr2	211179766	T	-	*MYL1*	frameshift deletion
chr3	71090482	-	C	*FOXP1*	frameshift insertion
chr4	123748299	-	C	*FGF2*	frameshift insertion
chr9	139405664	C	-	*NOTCH1*	frameshift deletion
chr10	88478558	-	C	*LDB3*	frameshift insertion
chr10	92679010	-	T	*ANKRD1*	frameshift insertion
chr10	99338053	G	-	*ANKRD2*	frameshift deletion
chr11	2869086	-	GG	*KCNQ1*	frameshift insertion
chr11	2869088	-	GC	*KCNQ1*	frameshift insertion
chr12	115109685	-	A	*TBX3*	frameshift insertion
chr12	124824739	-	GCCG	*NCOR2*	frameshift insertion
chr12	124885147	-	G	*NCOR2*	frameshift insertion
chr14	73664749	-	GG	*PSEN1*	frameshift insertion
chr16	3778897	-	C	*CREBBP*	frameshift insertion
chr20	6750839	-	G	*BMP2*	frameshift insertion
chr20	33334734	-	A	*NCOA6*	splice site

Table 5.3. Indels which were identified by both the BWA-Dindel and the BWA-GATK pipelines. All positions based on HG19 reference.

	Samples	Unique indels (Total)	Unique indels (Filtered)	Found in both	Unique to
BWA-GATK	133	2873	35	25	10
BWA-DINDEL	133	7223	317	25	292

Table 5.4. Number of indels identified by both pipelines and the number of indels unique to each.

5.4.5 Indel validations

Of the 25 indels identified by both pipelines 13 occurred in samples provided by the laboratory at Newcastle University and were therefore available for immediate validation. Of these, 6 were identified as true positives, 5 were found to be false positives, and 2 remain unconfirmed (Table 5.5). The remaining 12 indels occurred in samples that were not from Newcastle and I am still waiting for validation. Therefore, of the 13 indels that occurred in available samples currently 6 have been proven to be true positives (Table 5.5).

At the time of the writing of this thesis, the *MYL1* deletion (Chromosome 2, position 211179765) which is likely to be true because it is found within a variable T region and a single base deletion was validated one base pair position upstream from it, remains unconfirmed. There is also a recognised insertion of a “T”, rs71888939, which is one base position upstream from my validated insertion. I am also awaiting results for, the *NOTCH2* variant (Table 5.5) that is located in a region where the primers were not specific enough (i.e. they aligned to two regions of the chromosome), and will have to be redesigned. Furthermore, I identified a 4 base pair long insertion (-/GCCG) in the *NCOR2* gene. However, Sanger sequencing validated this variant as a reported 9bp (-/GCCGCTGCT) insertion, rs77661573 (Table 5.6).

Chromosome	Position	Reference	Variant	Gene	Validation notes
chr1	2235396	-	G	*SKI*	False
chr1	71418662	G	-	*PTGER3*	True
chr1	120612003	GG	-	*NOTCH2*	Unconfirmed
chr1	201334355	T	-	*TNNT2*	True
chr2	66739381	-	T	*MEIS1*	False
chr2	121747069	-	A	*GLI2*	False
chr2	211179765	-	T	*MYL1*	Possibly true (Variable T region)
chr2	211179766	T	-	*MYL1*	True
chr4	123748299	-	C	*FGF2*	False
chr9	139405664	C	-	*NOTCH1*	True
chr10	99338053	G	-	*ANKRD2*	True
chr11	2869088	-	GC	*KCNQ1*	False
chr12	124824739	-	GCCG	*NCOR2*	True, but validated as 9bp's long

Table 5.5. Number of indels identified by both analysis pipelines which were validated, or which were false positive calls. All positions based on HG19 reference.

Gene	Number of samples (%)	EVS5400	1000genomes	dbSNP	Control list
PTGER3	1(<1)	0	0.0041	0	0
TNNT2	1 (<1)	0	0	0	0
ANKRD2	1(<1)	0	0	0	0
NCOR2	16 (12)	0	0	rs77661573	0
MYL1	2 (1.5)	0	0	0	0
NOTCH1	1 (1)	0	0	0	0

Table 5.6. Variant frequency of the 6 validated indels which were identified using both pipelines. Sanger sequencing identified the *NCOR2* variant as a 9 base pair SNP, rs7761573.

5.5 Discussion

Next-generation sequencing identified 25 potentially disease causing indels in 133 patients suffering from congenital cardiovascular malformations. All of the identified indels were located in the coding regions of genes and resulted in frameshifts. As all of the targeted genes are suspected to be involved in cardiovascular malformations, these indels may be related to disease in these samples. However, further functional studies would still need to be conducted in order to identify the true consequences of these indel variants.

Of the 25 indels called by both pipelines, 13 were identified in samples from Newcastle, and therefore available for immediate validation. Although a validation rate of only 46% was achieved, this is consistent with, or better than other studies reported in the literature. For example Weber *et al.* (2002), only achieved confirmation rates of 14% and 58% for indels of 1bp and 2bp's long respectively. The 6 validated indels, all occurred in different genes, namely *PTGER3*, *ANKRD2*, *NCOR2*, *MYL1*, *NOTCH1*, and *TNNT2*. These genes will be discussed in more detail below. In addition, an indel was identified within the *DSC2* gene in 7 of thesamples. This indel was not present in the final list as it had a frequency of 1.4% in the control list. However, it is a previously reported insertion and has been proposed to influence cardiovascular development (Beffagna *et al.*, 2007; Syrris *et al.*, 2007; De Bortoli *et al.*, 2010; Gehmlich *et al.*, 2011). In addition, its population frequency has been estimated to be ~3% (Syrris *et al.*, 2007; De Bortoli *et al.*, 2010), and therefore its presence within the control list is not unexpected.

5.5.1 Insertion and deletion, MYL1

The Myosin Light Chain 1 (*MYL1*) gene is located at chromosomal position 2q33-q34, is 1052 base pairs long and comprises 7 exons (NCBI). I identified a single base insertion in one of the samples in *MYL1* that resulted in a frameshift of the first exon. It is located within a variable T region and was therefore difficult to validate (See figure 5.4). However, I also identified a single base deletion within the same variable T region in two additional samples. This deletion was only one base pair upstream from the previous insertion, and has been validated. It too results in a frameshift in exon 1 of the *MYL1* gene. There is also a SNP located one base pair position upstream from my deletion, rs71888939. This highlights the variability of this gene region, and the possible difficulties in correctly identifying the positions of indels in this region. Despite this variability, *MYL1* may still be of interest.

The first stage in the development of the heart involves the formation of myofibrils in the cardiomyocytes, which allow for heart contraction (England and Loughna, 2012). Once fully developed, heart muscle contraction is accomplished by myosin containing filaments pulling on filaments composed largely of actin (Timson, 2003). Muscle myosin is a hexamer consisting of two myosin heavy chains (MHC) and four myosin light chains (MYL) (Barton *et al.*, 1985; Timson, 2003; Rottbauer *et al.*, 2006; England and Loughna, 2012). MYL chains are encoded for by eight genes, one of which is *MYL1* (England and Loughna, 2012).

Genes encoding for MYL chains influence heart development, contraction and maintenance, and in model organisms, defects in these genes lead to hypertrophic cardiomyopathy (Shimada *et al.*, 2009). More specifically, the myosin light chains are involved in the regulation of heart contraction (Timson, 2003), and disruption to these genes will severely affect cardiac function (Huang *et al.*, 2003; Shimada *et al.*, 2009). The study by Huang *et al.* (2003) suggests that myosin light chain 2 (*MLC2a*) is very important in the development of the atrial myofibrillar apparatus. They developed mice mutants with a non-functioning *MLC2a* gene, all of which died before birth due to severe atrial malformations. Also, Rottbauer *et al.* (2006) demonstrated that the removal of *MLC2* function leads to a severe disruption of cardiac function in zebrafish. They found that atrial and ventricular cardiomyocytes in zebrafish in which *MLC2* function had been abolished, were unable to contract.

The insertion at position 211179765, chromosome 2 (c.2_3insA), was identified in a single individual displaying a pulmonary atresia with intact ventricular septum (PAIVS), while the deletion at position 211179766, chromosome 2 (c.1delA), was identified in two individuals displaying non compaction and PAIVS respectively. Both variants occur in a highly conserved region early on in exon 1 (Figure 5.4) of the gene, were predicted as potentially disease causing, and are not present in the EVS5400, 1000 genomes or in the dbSNP databases (Table 5.6). The variants are also not present in the control list, but the indel was present in 8 random controls sequenced by Dr. Elise Glen.

PAIVS is a rare congenital disorder, making up about 1% of all congenital cardiac disorders, and is characterised by a complete obstruction of blood flow from the right ventricle to the pulmonary trunk and left ventricle (Gutgesell, 1975; Trusler *et al.*, 1976; Ashburn *et al.*, 2004). It is a morphologically diverse malformation showing large variations in the anatomy of the right ventricle and coronary arteries (Bull *et al.*, 1982; Mi *et al.*, 2005).

Non compaction cardiomyopthies are heart muscle disorders which can arise in either children or adults and manifest as heart failure (Engberding *et al.*, 2010). As well as occurring as singular isolated cases, they can also occur within multiple members of families (Oechslin *et al.*, 2000; Ichida *et al.*, 2001; Engberding *et al.*, 2010). The genetic causes of non compaction cardiomyopathies are diverse and have been shown to be caused by various genes such as *MYH7*, *ACTC* and *TNNT2* (Klaassen *et al.*, 2008). Chapter 4 of my thesis also highlights other genes which have been shown to be responsible for some forms of cardiomyopathy.

The influence of myosin light chains on proper cardiac development, and subsequent function, make the *MYL1* a good candidate for the disease gene in these three individuals. However, the variability of this region will make it difficult to assess the true impact of these indels on disease.

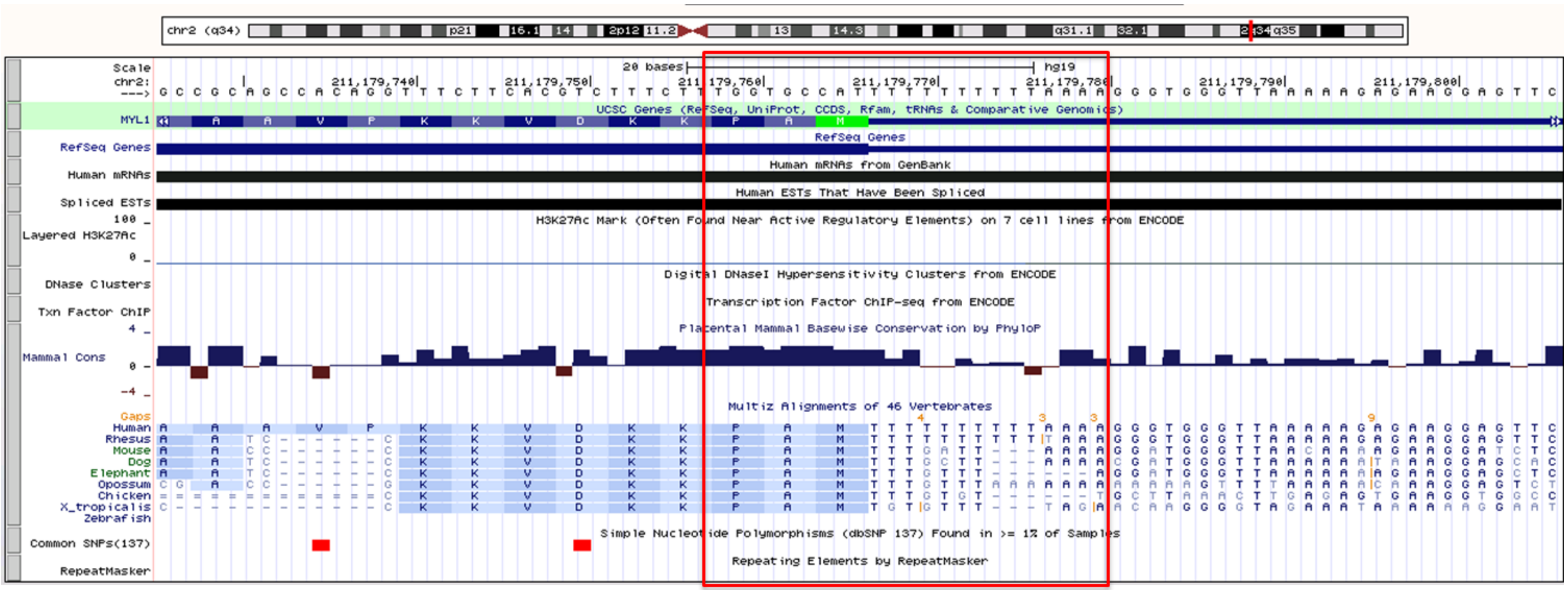

chr2 (q34)
p21
16.1
14
2p12
11.2
13
14.3
q31.1
32.1
q34
q35
Scale
chr2:
20 bases
hg19
211,179,740|
211,179,750|
211,179,760|
211,179,770|
211,179,780|
211,179,790|
211,179,800|
MYL1
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Figure 5.4. The position of the variable T region and the position of both indels which were identified in the *MYL1* gene (http://genome.ucsc.edu).

5.5.2 Deletion, NOTCH1

A single base deletion was identified in the *NOTCH1* gene. The *NOTCH1* gene is located at chromosomal position 9q34.3, is 9309 base pairs long and comprises 34 exons (NCBI). The deletion results in a frameshift in exon 16 of the gene (c.2527delG), and brings the reading frame forward by one position. It is in a highly conserved region (Figure 5.5), was predicted to be potentially disease causing, and not present in the EVS5400, 1000 genomes, or dbSNP databases (Table 5.6). It was also not present in the control list. However, I can not confirm whether it is a *de novo* variant as at the time of writing my thesis the fathers DNA was not available.

The *NOTCH1* gene forms part of the highly conserved Notch signalling pathway which is involved in cell-cell communications, in particular, it is an integral pathway involved in cell-fate determinations (Gordon *et al.*, 2008; de la Pompa and Epstein, 2012) and its signalling regulates organogenesis and cellular processes, including proliferation and apoptosis in mammals (MacGrogan *et al.*, 2011). NOTCH signalling is especially important for the formation of the heart, which requires the coordinated development of multiple parts (de la Pompa and Epstein, 2012). There are currently four recognised *NOTCH* genes which all play a very important role in proper cardiac development (High and Epstein, 2008; MacGrogan *et al.*, 2011; de la Pompa and Epstein, 2012).

Disruptions to the Notch signalling genes have been shown to cause various cardiovascular developmental disorders (Krantz *et al.*, 1999; Eldadah *et al.*, 2001; Garg *et al.*, 2005). MacGrogan *et al.* (2011) provide a good review of the role that NOTCH signalling plays in cardiac development. In particular during valve development, where the epithelial-mesenchyme transition, which is activated by endocardial Notch signalling, give rise to the valve primordial. For example, Garg *et al.* (2005) have demonstrated that mutations within the *NOTCH1* gene can lead to aortic valve disease. They analysed a multi-generational pedigree with 11 cases of congenital heart disease, 9 of which displayed aortic valve disease. Direct sequencing of the *NOTCH1* gene revealed a R1108X nonsense mutation present in all cases. This variant was not found in unaffected family members or in 1 136 control samples. There are also many mouse models supporting the influence of the Notch pathway genes on cardiovascular defects (Krebs *et al.*, 2000; Duarte *et al.*, 2004). For example, Krebs *et al.* (2000) developed both *NOTCH1*-deficient and *NOTCH4*-deficient mice. The resultant embryos had severe abnormalities in angiogenic vascular remodelling.

The insertion I identified (Chromosome 9, position 139405664) was found in a patient with mitral valve atresia. Mitral valve abnormalities are often complex, with severe forms occurring rarely (Remenyi and Gentles, 2012). The true incidence is difficult to assess as it is often classified with other congenital cardiac malformations (Summerell *et al.*, 1968). For example, in Marfan Syndrome where mitral valve disease is the leading cause of death in children suffering from this disorder (Ng *et al.*, 2004). However, the genetic causes of some forms of non-syndromic cardiac valve diseases have also been identified, such as in isolated, non-syndromic valvular dystrophy (Kyndt *et al.*, 2007). By sequencing the *FLNA* gene Kyndt *et al.* (2007) identified mutations in all of their 43 cases that were not present in unaffected family members or in 500 controls.

The influence of *NOTCH1* in various cardiac disorders involving abnormal valve development (Garg *et al.*, 2005; McKellar *et al.*, 2007; McBride *et al.*, 2008; Acharya *et al.*, 2011) make this gene a good candidate for disease in this patient.

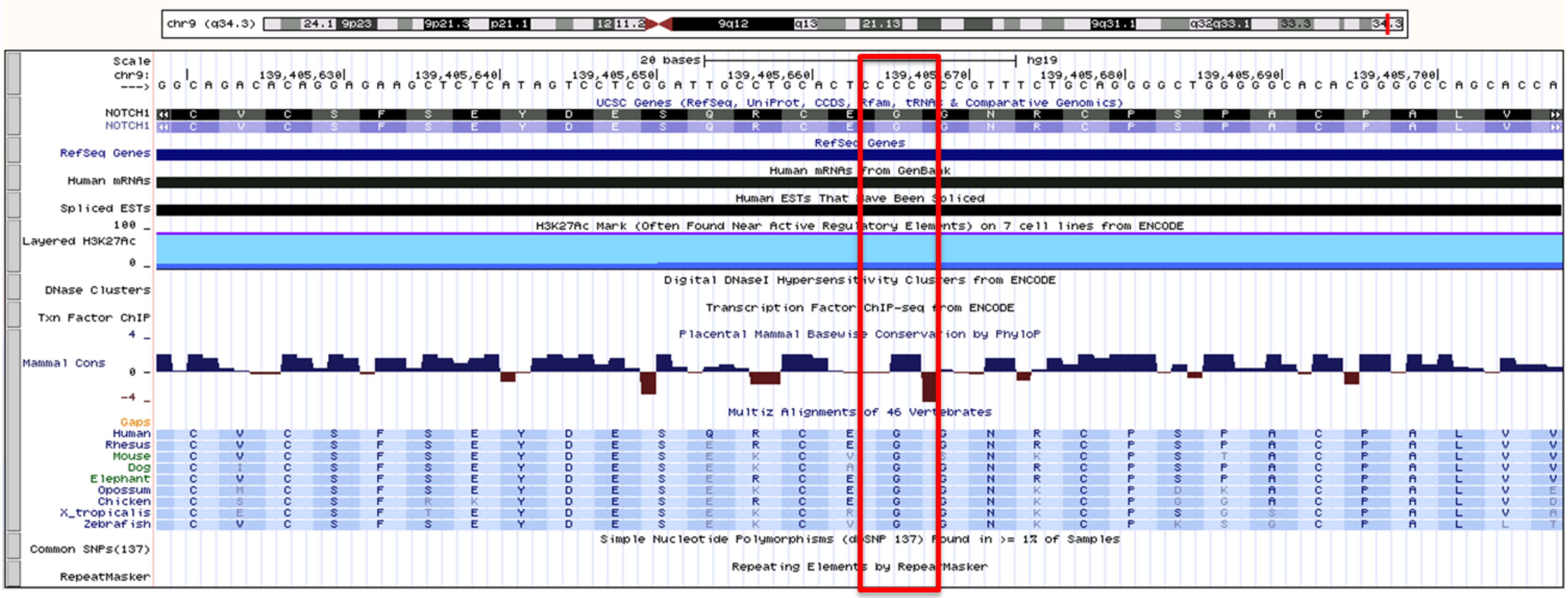

chr9 (q34.3)
24.1
9p23
9p21.3
p21.1
12
11.2
9q12
q13
21.13
9q31.1
q32
q33.1
33.3
34.3
Scale
chr9:
20 bases
hg19
139,405,630
139,405,640
139,405,650
139,405,660
139,405,670
139,405,680
139,405,690
139,405,700
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
NOTCH1
NOTCH1
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
4
Mammal Cons
0
-4
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 5.5. Indel which I identified in the *NOTCH1* gene (http://genome.ucsc.edu).

5.5.3 Deletion, TNNT2

The *TNNT2* gene is located at chromosomal position 1q32, is 1153 base pairs long, and comprises 17 exons (NCBI). I was able to identify and validate a single base deletion in the *TNNT2* gene (c.330delA). The deletion is in a conserved region and results in a frameshift in exon 9 of this gene (Figure 5.6). It was also predicted to be potentially disease causing by MutationTaster. *TNNT2* is part of the troponin protein complex that regulates the interaction of myosin and actin, thereby influencing the contraction of vertebrate striated muscle (Zot and Potter, 1987; Morimoto *et al.*, 2002; Huang *et al.*, 2009). The complex comprises three interacting proteins that stimulate contraction of the heart in response to the presence of Ca^{2+} (Morimoto *et al.*, 2002; Parmacek and Solaro, 2004; Huang *et al.*, 2009).

Deletions within cardiac troponin T genes, such as *TNNT2*, have been shown to cause various cardiomyopathies, such as dilated cardiomyopathy (Kamisago *et al.*, 2000; Morimoto *et al.*, 2002; Villard *et al.*, 2005) and hypertrophic cardiomyopathy (Forissier *et al.*, 1996; Marian and Roberts, 2001). For example, by sequencing the *TNNT2* gene, Forissier *et al.* (1996) were able to identify a Arg102Leu missense mutation found only in the four affected individuals. Mutation screening was performed using a single strand conformation polymorphism analysis and sequencing using the Biosystem 373A DNA sequencer. The variant was not present in the healthy family members, or in 92 healthy controls. Also, by sequencing various sarcomere protein encoding genes in a set of patients displaying dilated cardiomyopathy, Kamisago *et al.* (2000) were able to identify a three nucleotide deletion in troponin T in all the affected family members, but not in the healthy members of the family, or in 200 unrelated controls.

The deletion was only identified in a single patient displaying Ebsteins Anomaly (EA). EA is a complex congenital malformation characterised by a structural deformity of the tricuspid valve that results in a wide range of morphological and physiological changes (Correa-Villasenor *et al.*, 1994; Attenhofer Jost *et al.*, 2007). The malformation results in the abnormal flow of blood through the right side of the heart resulting in right ventricular dilation in about 60% of patients with EA (Attenhofer Jost *et al.*, 2007). Various genes have been found to cause EA, such as *GATA4* and *NKX2.4* (Digilio *et al.*, 2011). However, the enlargement of the heart chambers and decreased cardiac function through chamber enlargement seen in EA and the strong influence of *TNNT2* mutations on cardiomyopathies implicates this gene as possibly disease causing in this individual. Other possible genetic causes of EA have been discussed in more detail in chapter 4.

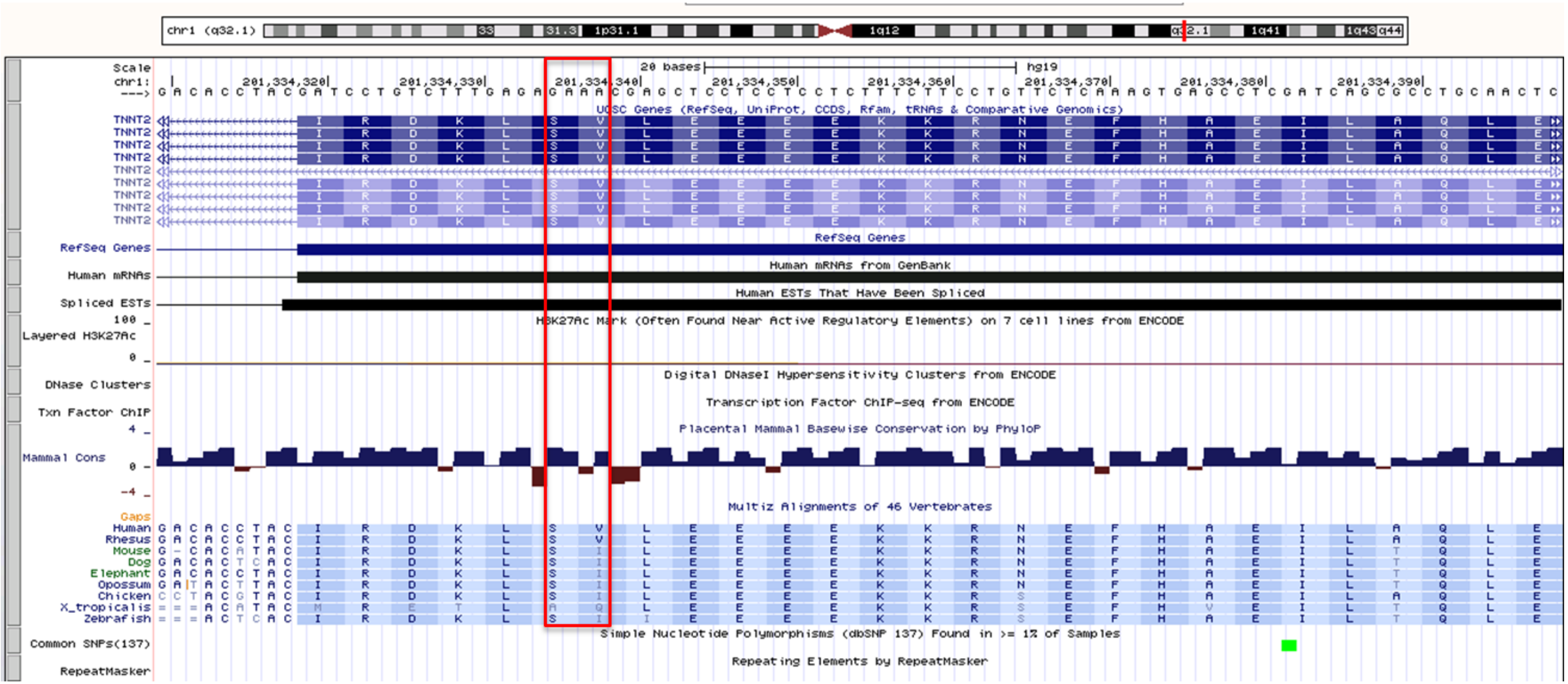

chr1 (q32.1)
33
31.3
1p31.1
1q12
q32.1
1q41
1q43
q44
Scale
chr1:
20 bases
hg19
201,334,320
201,334,330
201,334,340
201,334,350
201,334,360
201,334,370
201,334,380
201,334,390
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
TNNT2
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 5.6. Position of the identified indel in the *TNNT2* gene (http://genome.ucsc.edu).

5.5.4 Insertion, DSC2

The *DSC2* gene is found at chromosomal position 18q12.1, is 5257 base pairs long, and comprises 17 exons (NCBI). I identified a E896fsX900insertion in a conserved region of the *DSC2* gene (Figure 5.7), that was predicted to be potentially disease causing. It was found in 7 of the samples, but was not present in my final table because it had a frequency >1% in the control list. The presence of this insertion within the control list is expected as it is known to have a low level population frequency estimated to be ~3% (Syrris *et al.*, 2007; De Bortoli *et al.*, 2010). The insertion was identified in the final exon (Exon 17) of the gene, and affects the final 4 amino acids of the exon, truncating the protein. It was found in ~5% of the samples, and not present in the dbSNP database. However, it is present in the 1000 genomes project at a frequency of 1%. The difference in allele frequencies between the samples and the 1000 genomes is significant (*p*=0.00026). This specific variant has been previously reported in a number of studies assessing its potential as a cause of right ventricular cardiomyopathy (Syrris *et al.*, 2006; De Bortoli *et al.*, 2010; Gehmlich *et al.*, 2011).

A study by De Bortoli *et al.* (2010) found the variant in 5 of their 112 cardiomyopathy cases (allele frequency = 2.2%) but also in 6 out of 200 (allele frequency = 1.5%) of their healthy controls. This difference is not statistically significant and while not able to show an association of the insertion with disease, a functional analysis showed that the insertion altered the proper desmosomal localisation along cell boundaries. Also, alternative splicing produces two *DSC2* isoforms, *DSC2a* and *DSC2b* (Syrris *et al.*, 2006). The insertion only alters the *DSC2a* isoform and not the *DSC2b* isoform, and it is possible that *DSC2b* is compensating for the alteration in *DSC2a* in the control samples in which the insertion was identified (De Bortoli *et al.*, 2010). Syrris *et al.* (2006) state that due to the importance of desmocollins for cell adhesion, mutations in these genes would result in the decreased desmosome function and the possible detachment and death of cardiac myocytes, therefore negatively influencing cardiac development.

The patients in whom this variant was identified suffered from a range of cardiac abnormalities including EA, non compaction, PAIVS, right ventricular hypoplasia, double inlet left ventricle and hypoplastic left heart syndrome. Although, this variant has been reported to largely cause right ventricular cardiomyopathies (Syrris *et al.*, 2006; De Bortoli *et al.*, 2010; Gehmlich *et al.*, 2011), its potential effect on cardiac myocytes and therefore proper cardiac development could implicate it in a range of other cardiac developmental disorders.

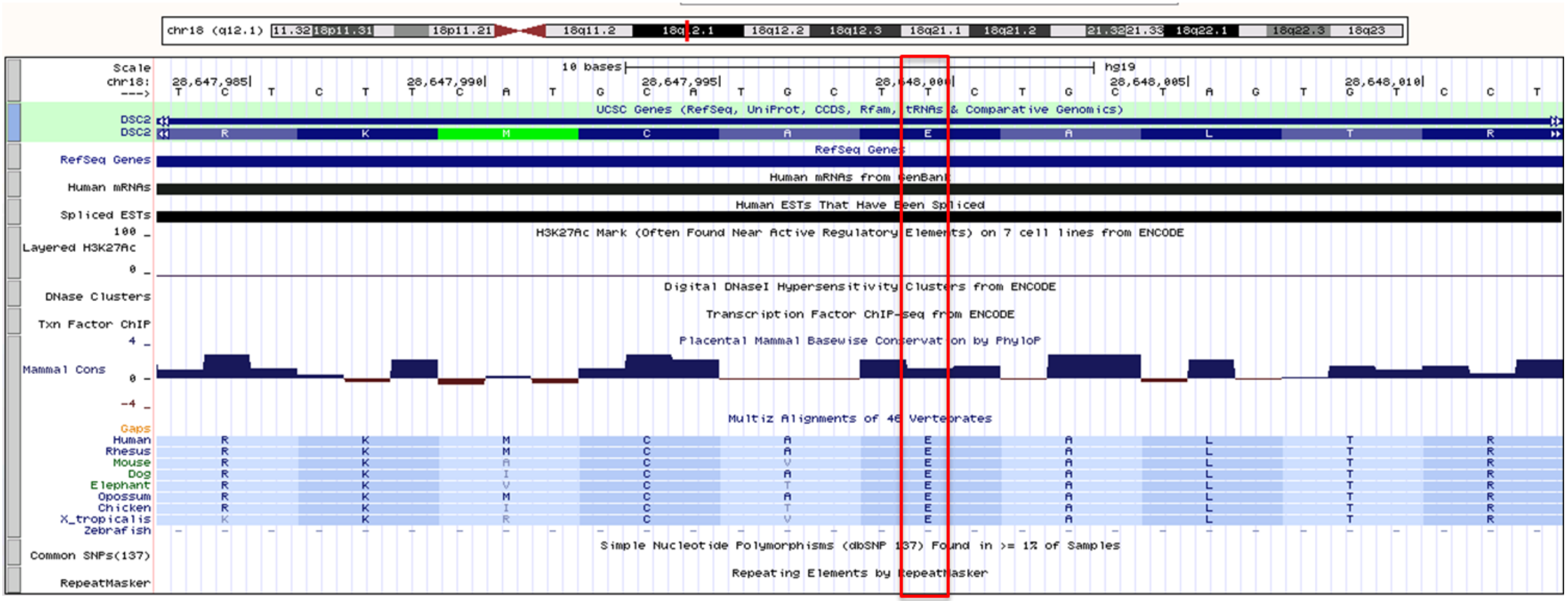

chr18 (q12.1)
11.32 18p11.31 18p11.21 18q11.2 18q12.1 18q12.2 18q12.3 18q21.1 18q21.2 21.32 21.33 18q22.1 18q22.3 18q23
Scale
chr18:
--->
10 bases
hg19
28,647,985
28,647,990
28,647,995
28,648,000
28,648,005
28,648,010
T C T C T T C A T G C A T G C T T C T G C T A G T G T C C T
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
DSC2
DSC2
R K M C A E A L T R
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Figure 5.7. Position and degree of conservation of the indel which I identified in the *DSC2* gene (http://genome.ucsc.edu).

5.5.5 Deletion, PTGER3

The *PTGER3* gene is a little known gene located at chromosomal position 1p31.2 (http://www.ncbi.nlm.nih). This gene is expressed in the heart and is part of the G-Protein coupled receptor family, and is one of the four receptors of Prostaglandin E2 which may be involved in several biological functions. The gene is involved in various pathways, some of which influence smooth muscle contraction and relaxation, figure 5.8. For example, *PTGER3* interacts with *KNG1* and *PTGER1* which both influence smooth muscle contraction (http://www.genecards.org).

I identified a c.1185delC deletion in the *PTGER3* gene in a single individual suffering from EA. The deletion is located in chromosome 1 at position 71418662 and results in a frameshift in the last exon of this gene (exon 4). It was predicted to be potentially disease causing and is in a region which is conserved in three other organisms (Figure 5.9), however, it is described as being intronic in all but one of the known transcripts (NM_198718). It is not found in the EVS or dbSNP databases, but the variant has been seen in the 1000 genomes at a frequency of below 1% (Table 5.6). Also, at the time of writing this thesis, the presence of this variant had not been validated in the parents, so I am unsure whether it is *de novo*.

Due to the limited information on this gene and the fact that the deletion occurs in the final exon of *PTGER3*, as well as it being described largely as an intronic region, it would be very difficult to identify the influence of this variant on disease in this patient.

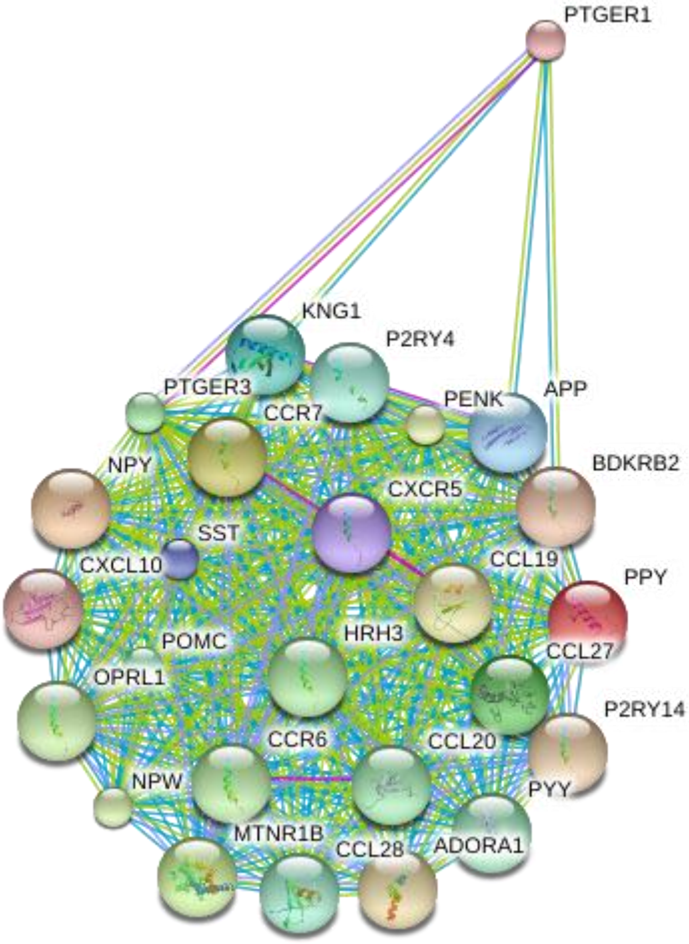


Figure 5.8. Genetic pathway interactions of the *PTGER3* gene. From http://www.genecards.org.

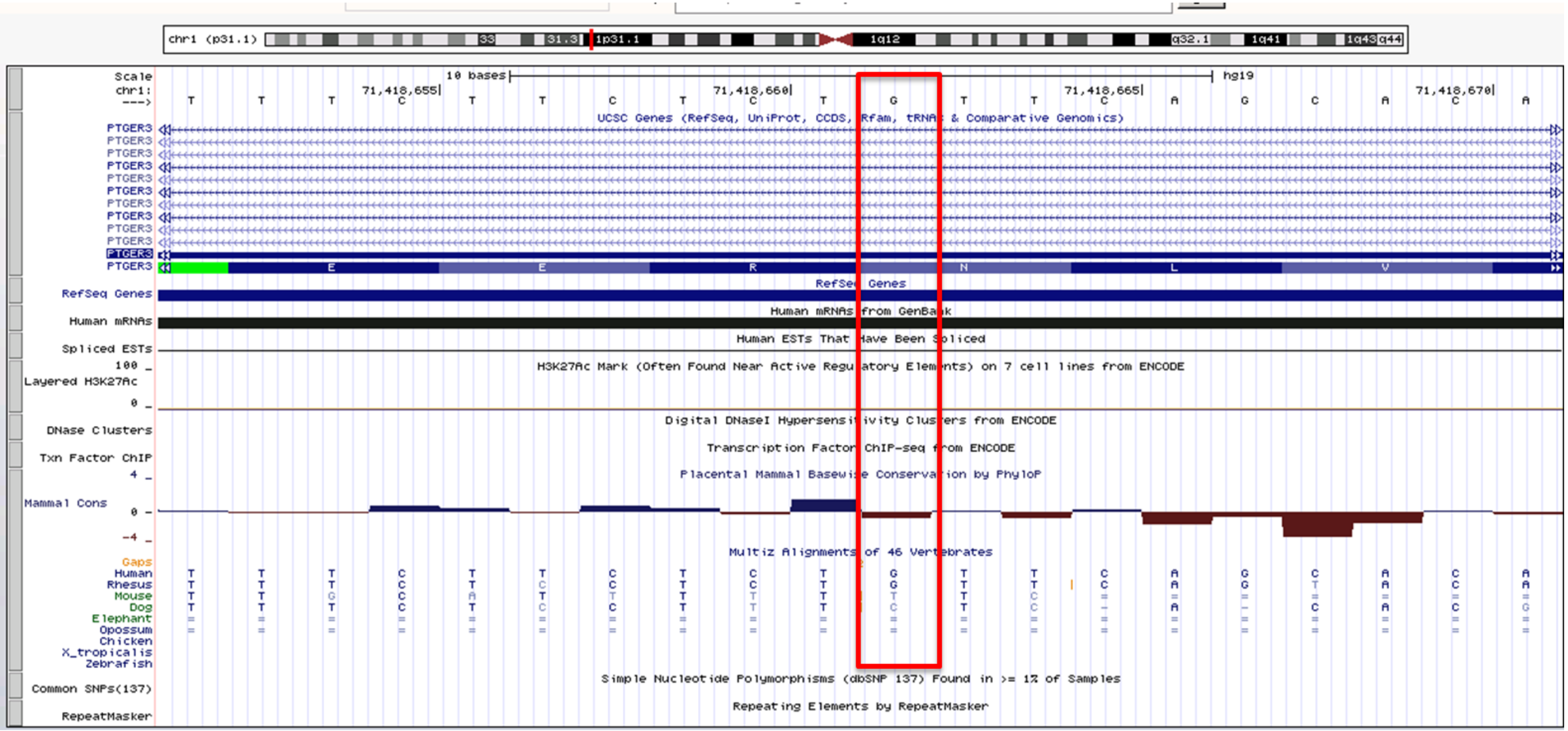
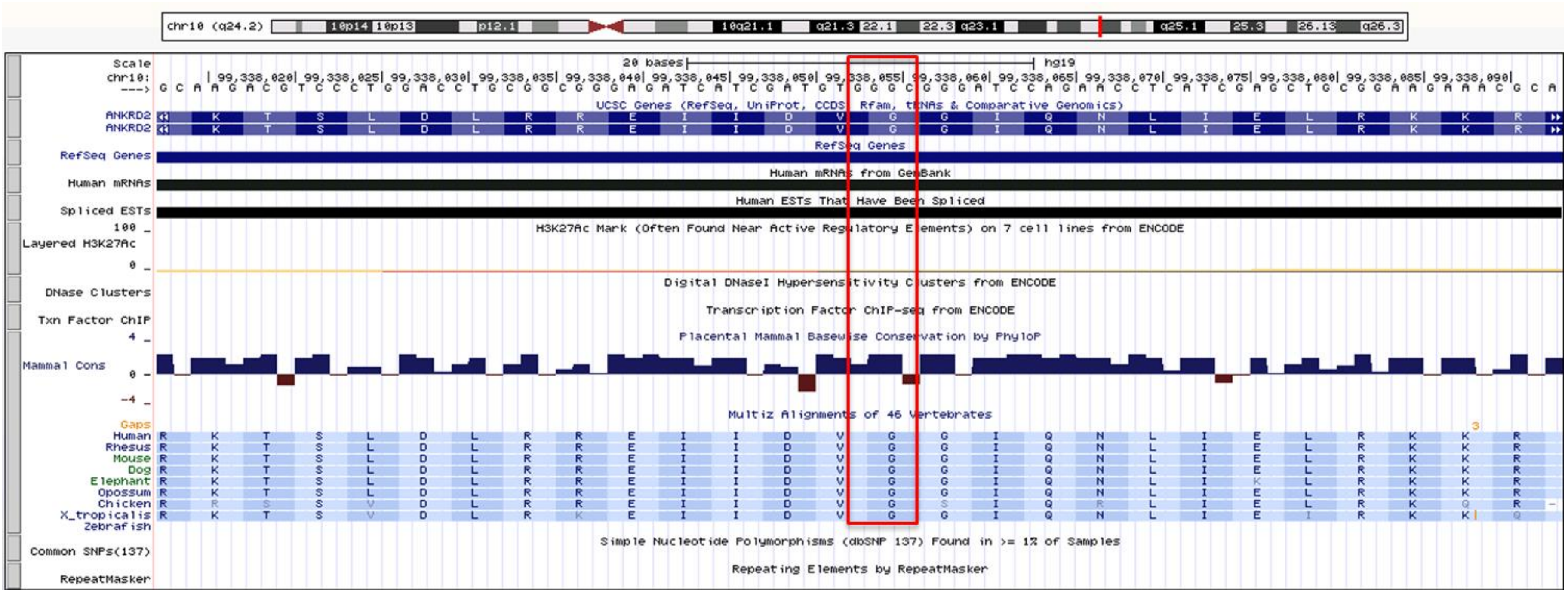

chr1 (p31.1)
33
31.3
1p31.1
1q12
q32.1
1q41
1q43
q44
Scale
chr1:
--->
10 bases
hg19
71,418,655
71,418,660
71,418,665
71,418,670
T T T C T T C T C T G T T C A G C A C A
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
PTGER3
E
E
R
N
L
V
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
4 _
Mammal Cons
0 -
-4 _
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 5.9. The deleted base from the *PTGER3* gene (http://genome.ucsc.edu).

5.5.6 Deletion, ANKRD2

The *ANKRD2* gene encodes the ANKRD2 protein which is one of the three members of the conserved muscle ankyrin repeat proteins that may be involved in muscle stress response pathways (Miller *et al.*, 2003). Human Ankrd2 is similar to proteins found in mice, rats and rabbits, but in all these organisms they are expressed predominantly in cardiac muscle, however in humans they are expressed predominantly in skeletal muscle (Pallavicini *et al.*, 2001; Belgrano *et al.*, 2011). It has been hypothesised, that although the human and mice proteins may be functionally related, they may well show specialisation for the tissues they are expressed in (Pallavicini *et al.*, 2001), therefore may not be active in the heart.

I identified a frameshift deletion in chromosome 10 (c.327delG) in the *ANKRD2* gene in a patient which suffers from double inlet left ventricle. At the time of writing my thesis, this variant had not been validated in the parents yet. The deletion was identified in a conserved region of the gene (Figure 5.10), and was predicted to be potentially disease causing. However, because of the features of Ankrd2 described above, I am unable to state whether this gene is a plausible candidate for disease in this patient.



chr10 (q24.2)
10p14 10p13 p12.1 10q21.1 q21.3 22.1 22.3 q23.1 q25.1 25.3 26.13 q26.3
Scale
20 bases
hg19
chr10:
99,338,020 99,338,025 99,338,030 99,338,035 99,338,040 99,338,045 99,338,050 99,338,055 99,338,060 99,338,065 99,338,070 99,338,075 99,338,080 99,338,085 99,338,090
--->
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
ANKRD2
ANKRD2
K T S L D L R R E I I D V G G I Q N L I E L R K K R
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4 _
Mammal Cons
0 -
-4 _
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Figure 5.10. The position and extent of conservation of the deleted base identified in the *ANRD2* gene (http://genome.ucsc.edu).

5.5.7 Insertion, NCOR2

The *NCOR2* gene encodes a nuclear receptor co-repressor responsible for the mediation of transcription silencing in certain genes. The gene encodes for a protein member of thyroid hormone and retinoic acid receptor associated co-repressors (http://www.genecards.org). This gene is part of the NOTCH signalling pathway (Figure 5.11), which as described earlier, is a very important pathway in heart development.

I identified a frameshift variant (c.5470_5471insCGGC) in exon 37 (Of 47) in this gene. The variant was identified in 16 patients suffering from a range of cardiac malformations including EA, noncompaction, tricuspid atresia, univentricular heart, pulmonary atresia with intact intraventriculal septum, and double inlet ventricle. The variant is located in a region conserved across a range of organisms (Figure 5.12), and was predicted to be potentially disease causing. However, the variant was validated as a 9bp insertion (/GCCGCTGCT) that is located in dbSNP (rs77661573), meaning that it is in fact not frameshifting. Also, rs77661573 is identified in the majority of people in the EVS database, and was validated using Sanger Sequencing (Dr. Elise Glen) in 7 random controls. The variant is also present in both parents. Therefore, this variant is likely not responsible for disease in these patients.

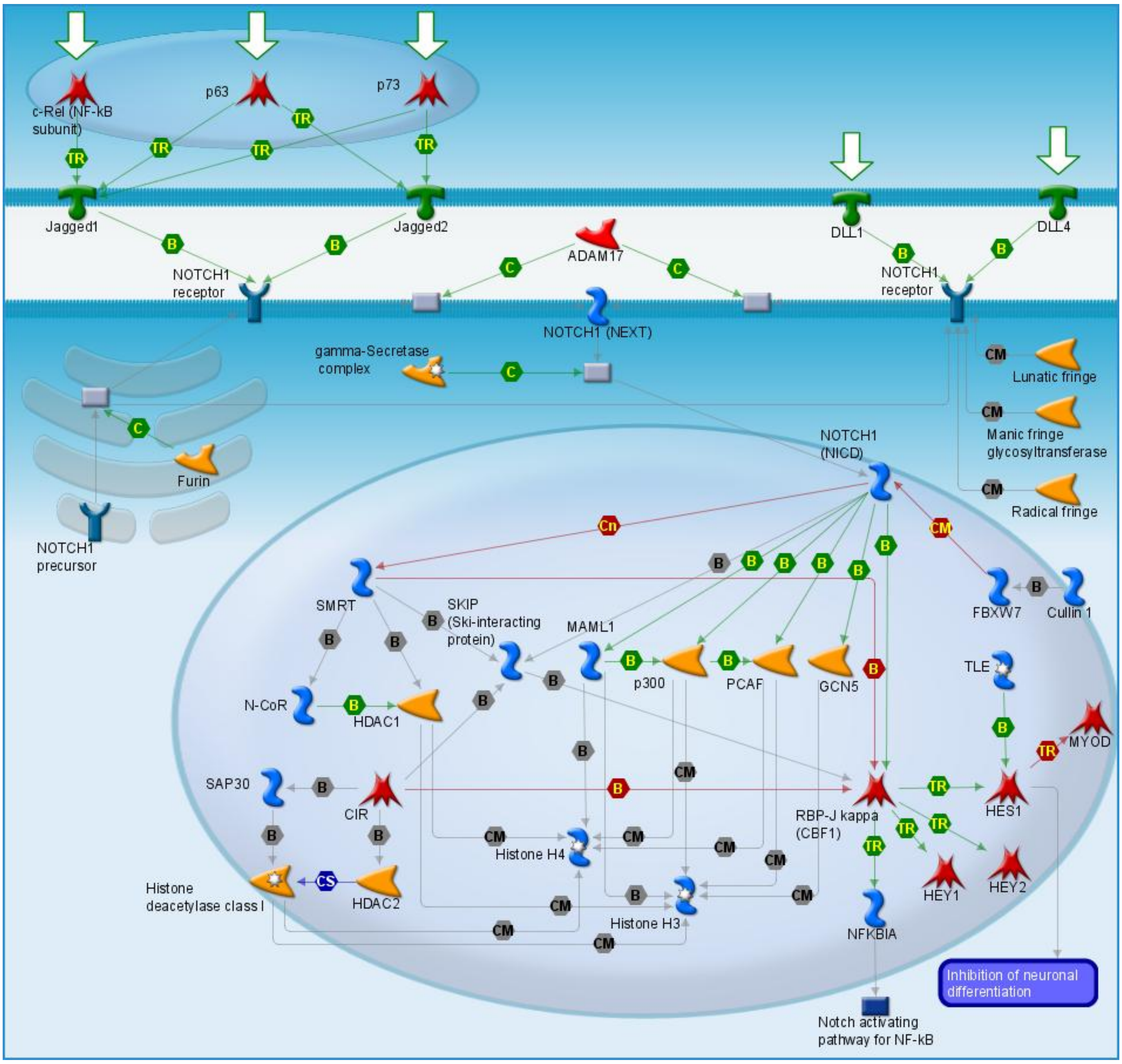


Figure 5.11. The NOTCH1 signalling pathway and the *NCOR* gene in this pathway (http://pathwaymaps.com).

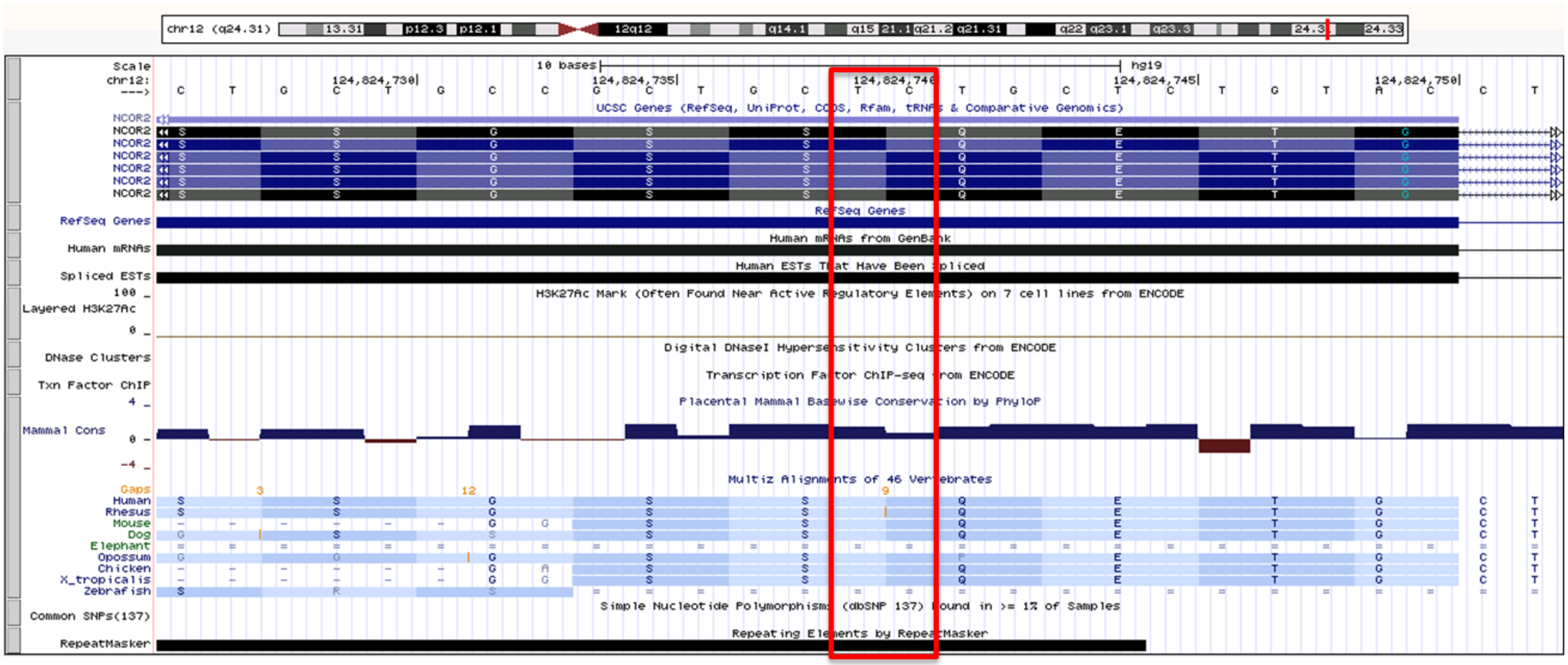

chr12 (q24.31)
13.31
p12.3
p12.1
12q12
q14.1
q15
21.1
q21.2
q21.31
q22
q23.1
q23.3
24.31
24.33
Scale
chr12:
--->
10 bases
hg19
124,824,730
124,824,735
124,824,740
124,824,745
124,824,750
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
NCOR2
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
100
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Mammal Cons
4
-4
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
3
12
9
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Figure 5.12. The inserted bases from the *NCOR2* gene (http://genome.ucsc.edu).

5.6 Conclusions

Using two different indel calling pipelines I was able to identify potentially disease causing variants in a group of 133 patients suffering from various congenital cardiovascular malformations, 46% of which were validated. All genes studied in this experiment had been selected for their influence on cardiac development and function, so it is possibly questionable to assert that certain genes out of the selected 403 are of greater interest than others *a posteriori*. However, it is noticeable that four of the genes harbouring indels are known causative genes for human cardiovascular disease, namely *MYL1*, *NOTCH1*, *TNNT2*, and *DSC2*. However it is important to note that these results do not infer causality, and despite their potential to influence disease in these patients, functional studies still need to be performed on these indels to assess their functional importance in cardiovascular development.

An important future direction for this work is to validate whether these indels are *de novo*. At the writing of this thesis, DNA was only available for some of the parents. Therefore, I could not assess whether all of the variants I identified were in fact *de novo*. Since these cases are assumed not to come from Mendelian families, Iassumed that in most cases, rare, *de novo* variants are responsible for disease in these patients. This would need to be assessed further.

The selection of the BWA-Dindel and BWA-GATK pipelines was done at the onset of the study based on current knowledge available in the literature; many studies had demonstrated that these were the most appropriate indel callers available at the time. Also, as this study progressed, so did the work in chapter 3 in which I designed a novel method for assessing the performance of NGS variant calls. Indeed, in my analysis BWA-Dindel and BWA-GATK did achieve the highest sensitivity values, in comparison to the other pipelines.

As the data chapters (Chapters 3, 4, and 5) of my thesis each contain in depth discussions, the summary discussion (Chapter 6) which follows, will recap the main findings of my thesis and highlight the areas which I think are important. It will also focus on the limitations and future directions of this work.

Chapter 6. Summary Discussion

6.1. Summary of findings

The aim of my PhD was to use NGS methods to identify rare, potentially disease causing variants involved in various diseases, particularly in CVM. My thesis took the form of three linked sub-projects, which developed concurrently with one another. In the first, I developed a novel approach to calculate the sensitivity and specificity of variant calls in NGS data using population SNP frequency information. This method allowed me to test the performance of various alignment and variant calling programmes. The NovoAlign-Samtools pipeline achieved the highest sensitivity and specificity values for identifying single nucleotide substitutions, and the BWA-Dindel and BWA-GATK pipelines achieved the highest sensitivity and specificity values for identifying indel variants.

The three pipelines mentioned above were used in the remaining two sub-projects of my PhD. In the first of these, I attempted to identify potentially disease causing variants in three families with disorders appearing to segregate in a Mendelian dominant fashion. These were atrioventricular septal defect (AVSD), dilated cardiomyopathy (DCM) and hereditary sclerosing poikiloderma (HSP). In the pedigrees where cases presented with DCM and HSP I was able to identify potentially disease causing variants in plausible candidate genes for disease. In the second sub-project, I used NGS to identify potentially disease causing and novel variants in a group of unrelated individuals suffering from various CVMs selected (by my colleagues in clinical cardiology) to involve either right or left ventricular hypoplasia. In these analyses I focused on indel variants, given the recognised challenges in correctly identifying these variants. In the 133 cases, evaluated for 403 candidate genes, I discovered 4 previously undescribed, frameshifting indels that, given the strong evolutionary constraints on such indels and the known consequences of variants in these genes, have a high *a priori* likelihood of being related to disease.

Since historically CVM has been a condition with a high early mortality, selective pressure on causative variants is likely to have been strong. Therefore, with regard to the adoption of NGS methodology in “sporadic” cases, it was a reasonable hypothesis when I commenced my project that common variants causing an increase in CVM risk might not exist, or be very few in number and of very small effect. I reasoned that risk alleles were more likely to be rare in the population, justifying a sequencing approach.

6.2 Limitations of this work

Limitations of the method proposed in chapter 3

My method for calculating sensitivity and specificity provided comparable results to methods requiring microarray data. Using this method, I was able to compare different variant calling pipelines and identify the best performing pipelines as those generating the highest sensitivity and specificity values. However, in the case of specificity, values are always close to 1. In fact, specificity values substantially lower than 1 would make NGS non-viable, as the number of false positives would be too large. Therefore, it is necessary to assess very small differences between pipelines in assigning a rank order. In instances where the number of variants is not high, such as was the case for indels, this assessment is difficult. This is not strictly a limitation of my method, but more a limitation of available data.

I made use of the HapMap and 1000 genomes databases to obtain population allele frequency information. I could increase the number of markers by including information from other databases. However, there is a large amount of overlap between the different databases, therefore this will not lead to a substantial increase in the number of markers.

An alternative means to increase the number of available markers for the calculation of specificity would be to include sites for which there are no reported variants and assume a low minor allele frequency for all of these sites. This would increase the number of sites used and improve the ability to estimate specificity. Additionally, since I was interested in using NGS to identify rare variants responsible for disease, it may be appropriate to use all the polymorphisms present in the coding regions, and possibly in regions outside of these. Although the results are not presented I did indeed try this by making use of all of the HapMap SNPs which were present in the Agilent whole exome targets, not just the on-target SNPs represented in the microarray. This resulted in a lower sensitivity being achieved for all pipelines, with the NovoAlign-Samtools pipeline achieving the highest sensitivity of ~85%. One possible reason for this drop in sensitivity is that the polymorphisms present on microarrays represent a selection based on criteria that include the probability of being efficiently typed. It is possible that NGS is accurate, or inaccurate, in exactly the same regions.

Results presented in chapter 3 based on my approach had indicated little difference in specificity between the GATK and Dindel pipelines. However, empirical evidence presented in chapter 5 indicated that there was in fact a difference in specificity values between the two pipelines with respect to indel calling. After filtering, the BWA-GATK pipeline only identified 35 indels, while the BWA-Dindel pipeline identified 317 indels. Variant calls from the BWA-Dindel pipeline achieved a laboratory validation rate below 30%, whereas when using the BWA-GATK pipeline a validation rate approaching 50% was achieved, indicating a significantly higher specificity with BWA-GATK that had not been detected using my method. The explanation for this apparent discrepancy is related to variation between the read coverage parameter selected in the two chapters. Due to the large number of windows produced when using the Bowtie2 aligner in combination with Dindel in chapter 3, a minimum threshold of 7 reads covering each indel variant had to be applied in all instances where Dindel was used. For all other indel callers tested in chapter 3, the default parameters were used, which only impose a minimum threshold of 1 read covering each indel. However, the BWA-Dindel pipeline in chapter 4 and 5 generated far fewer windows and as such I was able to make use of the default parameter sets (a minimum threshold of 1 read covering each indel). Due to the variability in the performance of different indel calling pipelines, to increase the confidence of the indel calls in chapter 5, I used the intersection of both the BWA-Dindel and BWA-GATK pipeline. This approach seemed to increase the validation rate and remove many of the false positive calls.

As well as the method I proposed, there are other performance measures which could be used. For example, a popular performance metric is the proportion of identified variants at sites known to be polymorphic. The known polymorphic sites are obtained from databases such as dbSNP and the 1000 genomes. Although this is a commonly used method to assess the performance of a variant calling pipeline, it will not allow for the identification of the best performing pipeline, because this estimate will be very high (>95%) in most cases. Also, the proportion of identified variants matching SNPs in public databases is likely to increase in the future as methods improve and as the number of SNPs in these databases increases. The increasing number of variants in these databases will mean that an even greater proportion of identified variants will match the polymorphic sites. However, this will also increase the power of my method for calculating sensitivity and specificity by providing a greater number of allele frequencies. Using the proportion of identified variants matching polymorphic sites in public databases as a performance measure may also bias results towards a particular pipeline. For example, using an analysis pipeline to identify variants, and then assessing its performance by matching the identified variants to a database which used the same (or a very similar) pipeline, will obviously lead to high degree of concordance.

An alternative performance metric to use could be by assessing base call quality, alignment quality or coverage. These measures are often used as performance metrics in the literature; however they only assess the performance of the base caller and aligner, not the entire pipeline. Conversely, my method provides a means of assessing the performance of an entire NGS analysis pipeline. However, all these methods could certainly be used in conjunction with one another to provide a complete overview of all aspects of the pipeline used in an NGS experiment. Further work could be undertaken to provide an appropriate framework to unify these methods and provide a suitable user interface for routine use; however, this was beyond the scope of the present work.

Limitations of causative variant identification in chapter 4

I was able to identify potentially disease causing variants in plausible candidate genes in both the DCM and the HSP pedigrees, but not in the AVSD pedigree. These genes were identified as potential candidates based on the current understanding of their function, as identified from databases such as OMIM. However, there are possibly many genes for which no functional information or influence on disorders displaying similar characteristics may be available, but which may still influence disease in these pedigrees. For example, in the case of the *FAM111B* gene I identified in the HSP pedigree. There is very little information available for this gene, and had it not been for the second pedigree which was identified by my collaborators in Nantes, France, I would not have identified this gene as a plausible candidate for disease in this family.

There are also several other potential reasons for failure to identify a likely causative variant in the AVSD pedigree. First, the causative variants may not have been captured by the exon capture kit, due to the various technical limitations of the method, or they may have been removed by one of the filtering steps (See chapter 3 and 4). In particular, the causative variant may well occur outside of the exonic regions.

Also, I relied on the MutationTaster programme to assess the potential pathogenicity of the variants. Many studies have highlighted the potential shortfalls of using such prediction programmes (See chapter 4 and 5). Indeed, had it not been for the identification of a second HSP pedigree, I would likely have removed the *FAM111B* gene as a plausible candidate as the variant was not predicted to be disease causing by MutationTaster.

Another possibility is that the variants causing disease in the AVSD pedigree are not necessarily distinct in all four affected individuals. One of the filtering steps I used involved selecting only variants common to all the affected individuals. This of course assumes that the presence of affected individuals with AVSD is as a result of them having inherited the same disease causing variant (Bamshad *et al.*, 2011). However, this may not necessarily be true and there is a chance that for at least one of the affected individuals in the pedigree, disease is as a result of a variant which was not inherited (Gilissen *et al.*, 2011).

Limitations of causative variant identification in chapter 5

Recent data from the 1000 Genomes and EVS projects indicate that frameshifting indels are not only much rarer than non-synonymous single nucleotide substitutions, but that they are evolutionarily younger, and therefore *a priori* have a higher likelihood of being disease-causing. However, due to the difficulties involved in identifying indels, if disease was indeed caused by such variants there is a chance that the causative variant was not identified in my analysis. Additionally, methods for identifying copy number variants using NGS approaches remain at an early stage of development; I did not attempt to study CNVs in my present work. Previous observations indicate that CNVs do indeed contribute to CVM risk (Greenway *et al.*, 2009; Soemedi *et al.*, 2012); however, in the analysis of sporadic patients, any that had been shown to have a potentially causative CNV (>1Mb) based on analyses done by others within the host lab on SNP chip data were removed from analysis.

One of the main limitations of this work is that of sample size. For instance, using simulated data, Kiezun *et al.* (2012) expect that over 10 000 exomes would be required to achieve sufficient statistical power to detect associations of rare variations with complex traits. This was highlighted in a study investigating the role of rare genetic variants in breast cancer, which targeted 507 genes implicated in DNA repair and sequenced these on an Illumina HiSeq2000 in 1 150 cases (Ruark *et al.*, 2013). The study identified 1 044 protein truncating variants, and stratified the genes based on the number of different, rare truncating variants present in the samples. The *PPM1D* gene was the most overrepresented in this regard. To further explore the role of *PPM1D* in breast and ovarian cancer they performed a large scale case-control replication experiment using 7 781 unrelated individuals with breast and/ovarian cancer and 5 861 controls. They identified protein truncating variants in 25 of the 7 781 cases, and only 1 of the 5 861 controls ($p=1.12 \times 10^{-5}$). This study highlights the large sample sizes which are likely to be required in such a case-control study. The relatively small sample size in this work represented all the patient resource from a multi-centre international collaboration, since the CVM phenotypes I studied are rare. The seriousness of the conditions and their rarity led to thehypothesis at the outset of this work that rare deleterious variants might be significantly over-represented in cases, even in a relatively small discovery cohort. The work was commenced before the bulk of the 1000 Genomes data (eg the paper of MacArthur *et al.* 2012) was released showing a large excess of rare variants in the population, due to bottlenecks and weak selection, compared to what would be predicted from previous simulation-based studies. The discovery from that data that each of us harbours about 100 strictly defined loss of function alleles and 20 fully inactivate genes, with more than 50 heterozygous OMIM alleles, clearly mandates much larger studies if rare variants are to be successfully identified.

Issues of causality

It is important to note that even though variants were identified in plausible candidate genes, this does not imply causality. For infrequently occurring or unique variants, laboratory validation for each variant may be the only route to establishing a causal relationship. For more frequently occurring variants, or aggregated variants within particular genes, causality may seem more likely where there is an overrepresentation of variants in cases when compared to controls (thus establishing association). However, causality still remains an experimental issue that was beyond the scope of this work. Of note, even among common variants identified by GWAS in various diseases, molecular mechanisms accounting for the associations have in general yet to be discovered.

6.3. Future directions

Bioinformatic challenges

As the method proposed in chapter 3 is simple to implement, it would be possible to integrate it into a variant caller. It could then be used in much the same way as the recalibration steps used by the GATK variant caller to improve variant calls. The benefit of such a procedure was demonstrated in chapter 3 where a base quality >20 resulted in a large drop in sensitivity, while a base quality of >30 only resulted in a small increase in specificity. The ideal analysis pipeline should maximise sensitivity and specificity by optimising analysis parameters such as the base quality threshold.

A further aspect which would have to be considered in any future work would be whether or not to carry out whole genome sequencing. In both the analysis of the family data and in the analysis of the unrelated HeartRepair samples a targeted sequencing approach was used. For the family data a whole exome targeting approach was used, and for the analysis of the unrelated samples, the exons of 403 genes were targeted. Possibly a better option would be to sequence the whole genome in all samples as this would capture more of the genetic information. However, a major problem with whole genome sequencing is that it will identify a great many variants for which very little/nothing is known. Also, the costs of whole genome sequencing are very high, for which a low coverage yield is obtained. Most whole genome sequencing experiments only achieve a coverage of ~10 fold. In chapter 3, I demonstrated that at a coverage of 10 fold, the sensitivity is only ~20%. However, whole genome sequencing does benefit from not having an enrichment step and the associated biases of enrichment.

Also, as mentioned earlier, it is quite possible that the diseases in these cases are not caused by single base substitutions or by indels. Of particular interest would be the identification of CNVs. At the time of writing my thesis, methods for identifying CNVs robustly using short-read NGS data had not been well established or tested, and I did not attempt to identify these types of genetic variants. However, large CNVs should be identified and assessed in any future analysis.

Future work for the variants identified in chapter 4

I was able to identify three genes of particular interest in the HSP family, namely *FAM111B*, *AGAP6*, and *CNTNAPSB*. These three genes are of great interest because they were mutated in both HSP pedigrees. Due to the rarity of HSP, identifying genes mutated in both the unrelated pedigrees should be prioritised in any future work. For *FAM111B*, my collaborators are currently performing skin biopsies on both pedigrees. *FAM111B* expression in skin fibroblasts will be compared, and the expression levels of "classic" fibrotic genes (eg *Collagen 1*, *TIMP-1*, *TGFB1*, *PDGF*) evaluated in fibroblasts from case and control patients. Although I will not be directly involved in these analyses, a similar approach can also be taken for the *AGAP6* and *CNTNAPSB* genes, where expression levels in cases and controls could be compared.

It is also possible to use exome data from a small number of samples to perform association tests between cases and controls. Samtools, for example, provides some functionality in this regard. However, these methods are not widely used, largely because they have not been thoroughly tested. However, methods such as these could provide extra supporting evidence for my results. Recently, methods have also been developed to identify CNV's using only a small number of exomes. The potential importance of CNV's in disease has been discussed throughout my PhD, and even though methods for CNV detection using only a small number of exomes are presently limited, it would be an important step in any future analysis on these samples.

Future work for the variants identified in chapter 5

With regards to the HeartRepair study of sporadic CVM cases, the sequencing will be repeated in cases and controls in a replication cohort on the genes of interest. Of particular interest are the *MYL1*, *NOTCH1*, *TNNT2*, and *DSC2* genes, because they are known causative genes for human cardiovascular disease. A similar approach has been performed for the single nucleotide polymorphisms which have been identified by collaborators in Belgium. Dr. Elise Glen (University of Newcastle) used the EVS database as a control database to obtain allele frequencies for the 403 genes used in the HeartRepair study. She then compared the allele frequencies from the 133 cases against this set of controls using a chi-square test. In this first round of analysis, the *NKX2.3* gene was found to be overrepresented for variants in the cases (chi2 = 1.17E-06). A similar approach could be used for the indel data to identify any genes showing an overrepresentation of variants; however, far larger numbers of samples would need to undergo targeted capture and sequencing in order to provide a sufficient number of indels to enable any such statistical comparison to be conducted. This could be used as a test case, and if any genes are shown to be overrepresented, additional cases and controls could be sequenced.

6.4 Concluding remarks

During my PhD, I developed a method to calculate the sensitivity and specificity of NGS variant calls, which unlike current methods, does not require microarray data as a reference. It is a fast and simple to use method which can be used to test the performance of an entire NGS analysis pipeline. Using a whole exome sequencing approach I was also able to identify potentially disease causing variants in three families displaying Mendelian disorders. Additionally, using targeted sequence data I was able to identify potentially disease causing indels in group of unrelated individuals suffering from various sporadic CVMs.

Appendices:

```
use warnings;
use Getopt::Long;
my $Index="";
my ($command1, $command2, $command3="";
GetOptions ("CWD:s"=>\$command1,"InputFile:s"=>\$command2,"SampleId:s"=>\$command3);
my $CWD=$command1;
my $Input=$command2;
my $Id=$command3;
my $CurrentDir=$CWD;
chomp $CurrentDir;
my @DirContent=`ls $CurrentDir`;
my $file="$CurrentDir/$Input";
my $ConvertedFile=ConvertFileForMutaionTaster($file,$Id);
my $snp2snippetOut=$CurrentDir."/snp2snippetResults_hg19_".$Id.".txt";
my $ErrorsRemoved=$CurrentDir."/snp2snippetResults_hg19_".$Id.".txt_ErrorsRemoved";
`perl snp2snippet.pl $Input -g ref.fa -t EnsemblTranscripts_37.R59.tsv > $snp2snippetOut`;
`perl snp2snippet_removeSBVerrors.pl --file $snp2snippetOut`;
`perl mutation_taster_batch_query.pl -i $ErrorsRemoved`;
`perl mutation_taster_results.pl`;

sub ConvertFileForMutaionTaster{
 my $file=$_[0];
 my $Id=$_[1];
open FILE, $file;
my $out=”Output.txt”;
open OUT, “>>$out”;
while(<FILE>){
  chomp $_;
  my @SplitLine=split('\t', $_);
  my $Position=$SplitLine[2];
  my $Id=$SplitLine[0]."_".$Position."_".$Position."_".$SplitLine[3]."_".$SplitLine[4];
  print OUT “$SplitLine[0]\t$Position\t$SplitLine[4]\t$Id\n";
}
close FILE;
close OUT;
return($Out);
exit;
```

Script 2.1 Script used to run MutationTaster.

```
#!/usr/bin/perl -w
my $currentdir=`pwd`;
chomp $currentdir;
my @dircontent=`ls $currentdir`;
my %insertion=();
my %deletion=();
my $in="Input.txt";
open IN, "$in";
while(<IN>){
    chomp $_;
    my @splitline=split('\t', $_);
    if(length $splitline[3]==1){
        my $inslength=length $splitline[4];
        my $adjinslength=$inslength-1;
        if(!exists $insertion{$adjinslength}){
            $insertion{$adjinslength}=1;
        }else{
            $insertion{$adjinslength}=$insertion{$adjinslength}+1;
        }
    }elsif(length $splitline[4]==1){
        my $dellength=length $splitline[3];
        my $adjdellength=$dellength-1;
        if(!exists $deletion{$adjdellength}){
            $deletion{$adjdellength}=1;
        }else{
            $deletion{$adjdellength}=$deletion{$adjdellength}+1;
        }
    }
}
close IN;
exit;
```

Script 2.2 Script used calculate the size distribution of insertions and deletions.

```
my $out="Output.txt";
open OUT, ">>$out";
my %gatk=();
my $gatk="gatkinput.txt";
open GATK, "$gatk";
while(<GATK>){
chomp $_;
if($_=~/chr\S+/){
my @splitline=split('\t', $_);
$match=”$splitline[0].$splitline[1].$splitline[2].$splitline[3].$splitline[4].$splitline[5].$splitline[6]”;
$gatk{$match}=1;
}}
close GATK;
my $counter=0;
my $dindel="bwainput.txt";
open DINDEL, "$dindel";
while(<DINDEL>){
chomp $_;
if($_=~/chr\S+/){
my @splitline=split('\t', $_);
$match=”$splitline[0]$splitline[1].$splitline[2].$splitline[3].$splitline[4].$splitline[5].$splitline[6]”;
if(exists $gatk{$match}){
my @splitvalues=split('_',$gatk{$match});
print OUT
"$splitline[0]\t$splitline[1]\t$splitline[2]\t$splitline[3]\t$splitline[4]\t$splitline[5]\t$splitline[6]\n";
}else{
$counter++;
}}}
close GATK;
close OUT;
print $counter;
exit;
```

Script 2.3 Script used to identify variant overlaps between the BWA-Dindel and BWA-GATK pipelines.

Gene name				
ACTC1	*NTF3*	*HOPX*	*hsa-mir-133a-1*	*ARSE*
ACTN2	*NTRK3*	*HSPB7*	*hsa-mir-133a-2*	*CLIC2*
ACVRL1 (ALK1)	*PDGFA*	*IGF1*	*hsa-mir-208a*	*SRY*
ADAM17	*PDLIM3*	*IGFBP3*	*hsa-mir-208b*	*TNNI1*
ADAM19	*PDPK1*	*ISL1*	*hsa-mir-15b*	*TNNI2*
ADRB1	*PHC1*	*ITGA11*	*hsa-mir-15a*	*NPTX1*
BMP2	*PITX2*	*ITGA4*	*hsa-mir-21*	*SPOCK3*
BMP4	*PKP2*	*ITGA7*	*mmu-mir-715*	
BMP10	*PLN*	*ITGB1BP2*	*mmu-mir-190*	
BMPR1A	*PPP3R1 (CNB1)*	*ITGB1BP3*	*mmu-mir-22*	
BMPR1B	*PRKAG2 (AMPK)*	*JAK2*	*mmu-mir-199a-1*	
BRAF	*PTPN11*	*JPH1*	*mmu-mir-15a*	
CAV3	*RXRA*	*LAMA2*	*mmu-mir-378*	
CFC1	*RYR2*	*LAMA5*	*mmu-mir-466j*	
CHD7	*SALL1*	*LBR*	*mmu-mir-17*	
CITED2	*SCN5A*	*LBX1*	*mmu-mir-23a*	
CREBBP	*SEMA3C*	*LIMK1*	*mmu-mir-143*	
CRELD1	*SGCB*	*MAP2K3*	*mmu-mir-23b*	
CRYAB	*SGCD*	*MAP2K6*	*mmu-mir-1186*	
CSRP3	*SGCG*	*MAP3K7IP1*	*hsa-mir-23b*	
DES	*SLC2A4*	*MAPK12*	*hsa-mir-27b*	
DSC2	*SLC6A6*	*MBNL1*	*hsa-mir-130a*	
DSG2	*SLC8A1*	*MBNL3*	*hsa-mir-106a*	
DSP	*SMYD1*	*MEF2A*	*hsa-mir-199a-1*	
EGFR	*SOX9*	*MEF2B*	*hsa-mir-199a-2*	
ELN	*SRF*	*MEF2D*	*hsa-mir-22*	
ERBB2	*TAZ*	*MET*	*hsa-mir-199b*	
ERBB3	*TBX1*	*MIB1*	*hsa-mir-202*	
ERBB4	*TBX2*	*MKL2*	*DGCR14*	
EVC	*TBX20*	*MRAS*	*CLTC*	
FBN1	*TBX3*	*MTPN*	*IL15*	
FGF2	*TBX5*	*MUSK*	*DVL2*	
FGF8	*TCAP*	*MYL1*	*SC5DL*	
FGF9	*TEAD1*	*MYL4*	*TFAP2B*	
FGFR1	*TGFB2*	*MYL5*	*CECR1*	
FGFR2	*TGFBR3*	*MYL6*	*CUGBP2*	
FOXC1	*TMEM43*	*MYL6B*	*PAX3*	
FOXC2	*TMPO*	*MYL9*	*DRAP1*	
FOXM1	*TNNC1*	*MYOCD*	*IGF2*	

Table 2.1 Continued

FOXP1	TNNI3	MYOD1	NODAL
GAB1	TNNT2	MYOG	CECR2
GATA4	TPM1	MYOM1	EXT1
GATA5	TXNRD2	MYOM2	SATB1
GATA6	UFD1L	PBRM1	NR2F2
GJA1	VCL	PGAM2	DRG2
GJA5	VEGFA	POU6F1	RAI1
GLA	WNT7b	PPP1R12A	IRX5
HAND1	ZFPM1	PPP1R12B	DGCR2
HAND2	ZFPM2 (FOG2)	PPP3CA	DVL1
HBEGF	ZIC3	PPP3CB	ENG
HDAC2	SMAD6	PRDM6	KCNJ2
HDAC5	ROCK1	PRKAR1A	NR2C2
HDAC7A	WNT5a	PRKCA	Irx3
HDAC9	ISL1	PRKDC	PLXNA2
HEY1	FOXA2	PRKG1	Hoxb2
HEY2	BOP1	PSEN1	HTR2B
HHEX	ANK2	PTGER3	SHH
HIRA	ANKRD2	PTGER2	KCNQ1
HOPX (HOP)	ANKRD1	PTPRJ	KCNE1
HOXA3	BARX2	RAB3GAP2	NFATC4
HRAS	BARX1	SHOX2	CITED1
IDUA	CASQ2	SMYD1	NFATC1
IGF1R	CAV2	SOX15	NSD1
INSR	CNBP	SOX2	FKBP6
IRX4	CTF1	SOX6	TBL2
JAG1	CXADR	TBX18	NDN
JAG2	DNER	TEAD1	UBE3A
JUN	DVL3	WNT3A	PRKCZ
JUP	EDN2	SIRT2	EXO1
KCNA5	DYRK1B	SMPX	SH3YL1
KRAS	EFEMP2	SMTN	SEPT2
LAMP2	EGLN1	SSPN	CHL1
LDB3	EGR3	TMOD4	NCBP2
LEFTY1	ELN	WNT4	IRF2
LEFTY2	EMD	ZEB2	LMBR1
LMNA	EVC2	DPF3	MAML1
MAP2K1 (MEK1)	FBLN5	TLL1	MAFK
MAP2K2 (MEK2)	FGF12	SOX4	SMARCA1
MAPK14	FGF19	MEIS1	CACNA1B
MEF2C	FGF2	ACVR2B	GTPBP4

Table 2.1 Continued

MESP1	*FGF6*	*ZYX*	*BCCIP*
MYBPC3	*FGF9*	*NCAM1*	*YY1AP1*
MYH6	*FHL1*	*SKI*	*NINJ2*
MYL2	*FHL3*	*FOXO3*	*CHFR*
MYL3	*FKRP*	*SIRT1*	*RAN*
MYL7	*FKTN*	*HES1*	*CDC16*
MYLK2	*FLNC*	*LBH*	*PCSK6*
MYOZ2	*FOXH1*	*LRRC20*	*OCA2*
NCOA6	*FOXK1*	*TWIST1*	*PIGQ*
NCOR2	*FOXL2*	*PIAS1*	*GALNS*
NF1	*FOXO4*	*MSX1*	*RPA1*
NFATC3	*FOXP1*	*MYLK3*	*ADCYAP1*
NKX2-3	*GLI2*	*MYH7*	*ADNP2*
NKX2-5	*GTF2I*	*MAP3K7IP2*	*FSTL3*
NOS3	*GTF2IRD1*	*ADAM17*	*PEG3AS*
NOTCH1	*HDAC4*	*PROX1*	*BIRC7*
NOTCH2	*HDAC5*	*hsa-mir-1-2*	*PRMT2*
NPPA	*HDAC7*	*hsa-mir-1-1*	*NCAM2*
NRG1	*HDAC9*	*hsa-mir-133b*	*TYMP*

Table 2.1. Lists the genes used in the HeartRepair study.

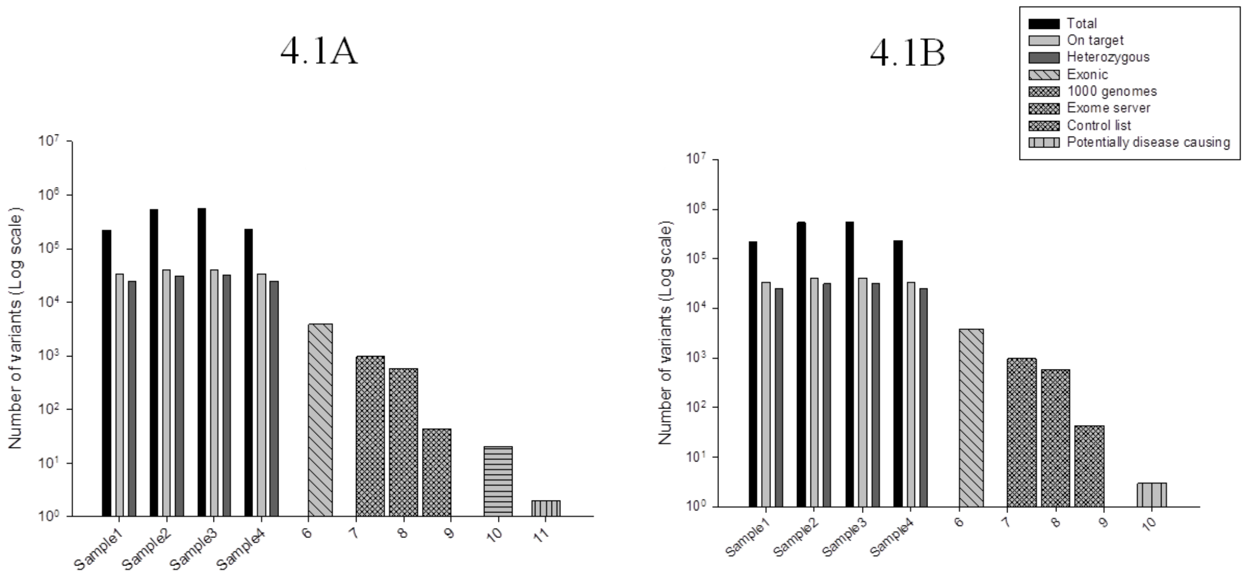


Figure 4.1. Influence of the different filtering steps on variant numbers in the DCM cases. 4.1A = Filtering set A, 4.1B = Filtering set B.

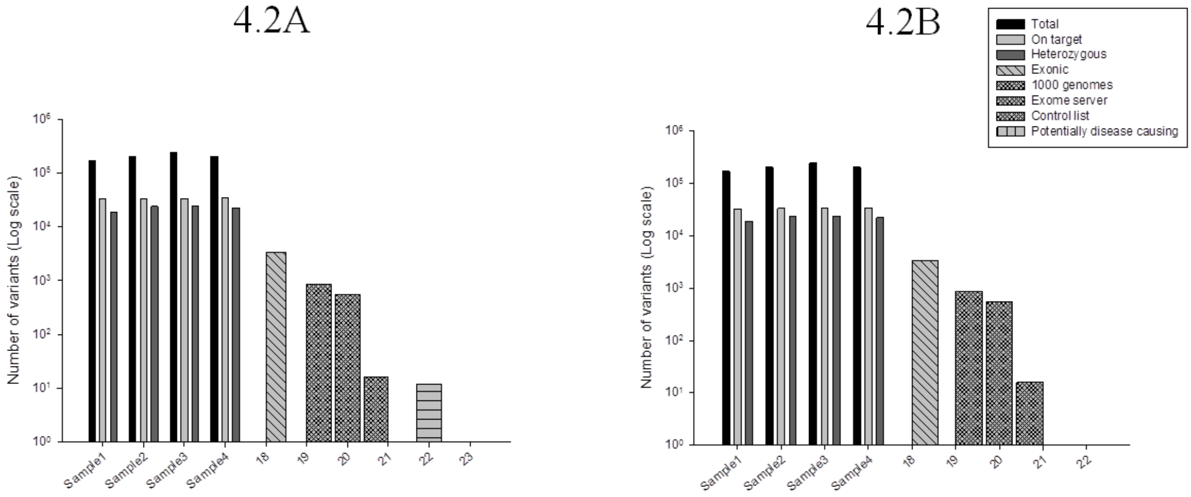


Figure 4.2. Influence of the different filtering steps on variant numbers in the AVSD cases. 4.2A = Filtering set A, 4.2B = Filtering set B.

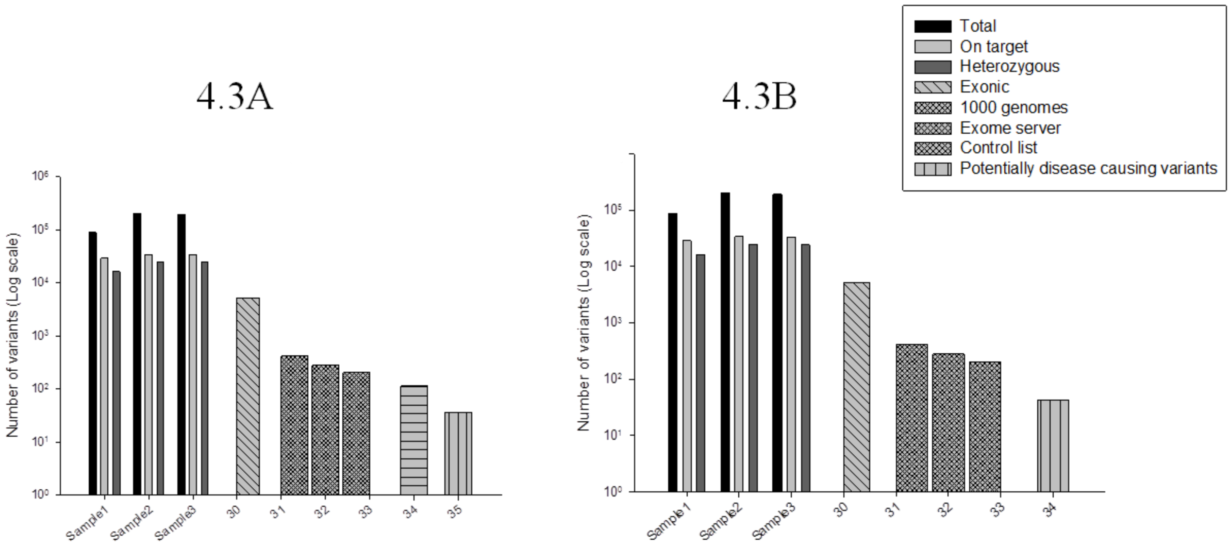


Figure 4.3. Influence of the different filtering steps on variant numbers in the HSP cases. 4.3A = Filtering set A, 4.3B = Filtering set B.

Sample	%target covered >20fold	%target covered >10fold	%target covered >5fold	%target covered >1fold
Sample1	82.64	89.88	93.42	97.25
Sample2	94.19	96.10	97.22	98.61
Sample3	95.78	97.02	97.81	98.78
Sample4	92.47	95.04	96.60	98.25
Sample5	94.92	96.29	97.30	98.53
Sample6	87.24	91.98	94.77	97.75
Sample7	92.01	94.98	96.66	98.46
Sample8	89.57	93.43	95.59	97.91
Sample9	92.54	95.07	96.57	98.35
Sample10	89.53	93.23	95.46	97.97
Sample11	91.69	94.81	96.48	98.36
Sample12	89.21	93.01	95.18	97.71
Sample13	89.96	93.69	95.91	98.18
Sample14	89.46	93.33	95.56	97.98
Sample15	88.27	92.41	95.01	97.90
Sample16	91.63	94.40	96.12	98.11
Sample17	92.45	95.03	96.61	98.38
Sample18	90.99	94.23	96.08	98.16
Sample19	45.19	77.67	88.47	95.88
Sample20	92.39	95.14	96.71	98.46
Sample21	93.96	95.95	97.21	98.54
Sample22	88.90	93.06	95.31	97.90
Sample23	93.74	95.86	97.15	98.57
Sample24	89.96	93.59	95.72	97.99
Sample25	88.60	92.91	95.26	97.86
Sample26	89.36	93.52	95.77	98.08
Sample27	85.96	91.68	94.74	97.83
Sample28	85.42	91.18	94.22	97.55
Sample29	85.39	91.24	94.35	97.63
Sample30	89.46	93.50	95.72	98.06
Sample31	94.33	96.20	97.35	98.59
Sample32	88.46	93.09	95.56	98.08
Sample33	84.48	91.64	94.84	97.47
Sample34	95.18	96.35	97.17	98.27
Sample35	89.84	93.85	96.08	98.38
Sample36	85.22	92.61	95.61	98.14
Sample37	93.93	95.73	96.96	98.41
Sample38	69.93	86.57	92.86	97.53
Sample39	91.05	94.32	96.23	98.35
Sample40	91.42	94.43	96.32	98.32
Sample41	95.03	96.61	97.65	98.69
Sample42	83.39	90.54	94.03	97.50
Sample43	86.76	93.37	96.41	98.50
Sample44	83.60	90.97	94.29	97.64

Table 5.1 Continued

Sample45	93.25	95.44	96.88	98.43
Sample46	93.80	95.69	97.01	98.42
Sample47	88.18	92.78	95.25	97.97
Sample48	93.65	95.68	96.99	98.50
Sample49	94.24	96.09	97.31	98.60
Sample50	83.06	89.67	93.38	97.05
Sample51	86.74	92.16	94.96	97.78
Sample52	83.81	91.85	95.42	98.13
Sample53	94.67	96.42	97.42	98.65
Sample54	87.34	92.19	94.90	97.74
Sample55	84.01	89.81	93.30	97.15
Sample56	83.14	90.70	94.18	97.62
Sample57	89.85	93.58	95.71	98.24
Sample58	88.08	94.14	96.83	98.62
Sample59	94.00	95.96	97.13	98.50
Sample60	91.91	94.68	96.31	98.14
Sample61	94.25	96.00	97.24	98.66
Sample62	92.04	94.80	96.33	98.26
Sample63	96.41	97.39	98.00	98.73
Sample64	93.10	95.45	96.85	98.30
Sample65	85.07	90.52	93.82	97.16
Sample66	95.30	96.72	97.64	98.62
Sample67	91.50	94.25	95.84	97.79
Sample68	90.36	93.61	95.67	97.99
Sample69	88.47	92.60	94.89	97.53
Sample70	90.82	93.82	95.61	97.92
Sample71	93.47	95.44	96.70	98.28
Sample72	94.10	95.91	97.09	98.49
Sample73	91.63	94.38	96.14	98.06
Sample74	90.51	94.17	95.95	98.08
Sample75	92.89	95.11	96.45	98.21
Sample76	93.07	95.40	96.82	98.38
Sample77	86.47	93.32	96.46	98.55
Sample78	91.64	94.61	96.32	98.23
Sample79	86.26	91.01	93.96	97.33
Sample80	89.51	93.48	95.69	98.15
Sample81	90.57	93.82	95.85	98.08
Sample82	94.21	96.03	97.24	98.60
Sample83	85.36	90.59	93.76	97.43
Sample84	88.93	92.52	94.92	97.74
Sample85	92.45	95.10	96.66	98.43
Sample86	83.20	89.43	92.73	96.44
Sample87	89.38	93.88	96.03	98.03
Sample88	77.12	87.62	92.87	97.13
Sample89	90.65	94.52	96.44	98.27

Table 5.1 Continued

Sample90	92.34	95.13	96.65	98.27
Sample91	87.35	92.24	94.97	97.71
Sample92	91.98	94.68	96.31	98.11
Sample93	81.06	89.59	93.37	97.10
Sample94	92.01	94.85	96.54	98.27
Sample95	90.92	94.06	95.76	97.72
Sample96	92.92	95.40	96.88	98.49
Sample97	91.38	94.11	95.67	97.51
Sample98	87.26	92.04	94.69	97.57
Sample99	91.54	94.40	96.12	97.94
Sample100	88.44	92.43	94.91	97.74
Sample101	80.52	88.19	92.15	96.22
Sample102	91.13	94.17	96.01	98.06
Sample103	82.59	89.45	93.08	96.87
Sample104	86.31	91.73	94.57	97.63
Sample105	93.65	95.35	96.55	97.92
Sample106	74.59	86.19	91.45	96.23
Sample107	80.94	88.74	92.68	96.72
Sample108	91.91	94.60	96.25	97.99
Sample109	88.88	92.77	95.17	97.92
Sample110	90.41	93.72	95.73	97.78
Sample111	94.75	96.33	97.44	98.60
Sample112	90.57	93.76	95.63	97.67
Sample113	89.93	93.20	95.16	97.36
Sample114	90.59	94.14	96.17	98.25
Sample115	87.03	91.60	94.49	97.61
Sample116	85.65	92.09	95.01	97.89
Sample117	88.24	92.76	95.28	97.95
Sample118	83.21	90.73	94.15	97.38
Sample119	91.61	94.48	96.23	98.21
Sample120	94.45	96.22	97.41	98.68
Sample121	89.34	93.53	95.73	98.17
Sample122	87.14	92.13	94.76	97.69
Sample123	90.53	93.99	95.94	98.07
Sample124	87.58	92.63	95.24	98.00
Sample125	91.21	94.46	96.25	98.36
Sample126	92.69	95.26	96.74	98.50
Sample127	91.14	94.48	96.33	98.35
Sample128	85.51	90.90	93.93	97.31
Sample129	90.47	93.85	95.89	98.18
Sample130	89.81	93.62	95.75	98.09
Sample131	80.88	89.75	93.65	97.33
Sample132	88.38	92.73	95.10	97.80
Sample133	88.49	92.93	95.41	97.89

Table 5.1. Target base coverage of HeartRepair samples.

Chrom.	Position	Ref.	Variant	Gene	EVS5400	1000 Genomes	dbsnp135	control list
chr11	65688382	G	-	*DRAP1*	0	0		0
chr9	140807675	A	-	*CACNA1B*	0	0		0
chr19	19258537	A	-	*MEF2B,MEF2BNB-MEF2B*	0	0		0
chr17	17697123	GC	-	*RAI1*	0	0		0
chr16	86601682	G	-	*FOXC2*	0	0		0
chr9	140917757	C	-	*CACNA1B*	0	0		0
chr5	156021946	A	-	*SGCD*	0	0		0
chr15	68606119	G	-	*ITGA11*	0	0		0
chr20	60905911	G	-	*LAMA5*	0	0		0
chr1	120612003	GG	-	*NOTCH2*	0	0		0
chr16	3778372	C	-	*CREBBP*	0	0		0
chr5	122515971	A	-	*PRDM6*	0	0		0
chr2	240002822	T	-	*HDAC4*	0	0		0
chr6	1611589	C	-	*FOXC1*	0	0		0
chr2	220290413	G	-	*DES*	0	0		0
chr16	88601159	C	-	*ZFPM1*	0	0		0
chr22	19754108	G	-	*TBX1*	0	0		0
chr1	201334355	T	-	*TNNT2*	0	0		0
chr12	9091915	G	-	*PHC1*	0	0		0
chr3	38622801	G	-	*SCN5A*	0	0		0
chr8	38271271	C	-	*FGFR1*	0	0		0
chr12	124820088	T	-	*NCOR2*	0	0		0
chr10	99338053	G	-	*ANKRD2*	0	0		0
chr5	122425851	CA	-	*PRDM6*	0	0		0
chr3	157823620	C	-	*SHOX2*	0	0		0
chr1	202407190	T	-	*PPP1R12B*	0	0		0
chr1	156106799	A	-	*LMNA*	0	0		0
chr10	72061238	T	-	*LRRC20*	0	0		0
chr12	5154544	T	-	*KCNA5*	0	0		0
chr9	140918091	C	-	*CACNA1B*	0	0		0
chr6	43139816	G	-	*SRF*	0	0		0
chr10	99337655	G	-	*ANKRD2*	0	0		0
chr7	74149837	AAGA	-	*GTF2I*	0	0		0
chr2	211179766	T	-	*MYL1*	0	0		0
chr19	4102407	A	-	*MAP2K2*	0	0		0
chr14	23895248	T	-	*MYH7*	0	0		0
chr1	155630185	C	-	*YY1AP1*	0	0		0
chr12	124831127	C	-	*NCOR2*	0	0		0
chr19	47259864	T	-	*FKRP*	0	0		0
chr6	50791258	TA	-	*TFAP2B*	0	0		0
chr1	220379328	T	-	*RAB3GAP2*	0	0		0
chr9	141013169	G	-	*CACNA1B*	0	0		0
chr8	145699791	A	-	*FOXH1*	0	0		0

Table 5.2 Continued

chr16	624455	G	-	*PIGQ*	0	0		0
chr20	60895929	C	-	*LAMA5*	0	0		0
chr4	102001782	A	-	*PPP3CA*	0	0		0
chr11	2869100	A	-	*KCNQ1*	0	0		0
chr9	139396296	G	-	*NOTCH1*	0	0		0
chr9	139405664	C	-	*NOTCH1*	0	0		0
chr12	9085217	C	-	*PHC1*	0	0		0
chr7	73521406	G	-	*LIMK1*	0	0		0
chr1	156106096	A	-	*LMNA*	0	0		0
chr5	88056864	C	-	*MEF2C*	0	0		0
chr11	2154242	G	-	*IGF2*	0	0		0
chr12	56491631	A	-	*ERBB3*	0	0		0
chr1	202533599	A	-	*PPP1R12B*	0	0		0
chrX	131518723	G	-	*MBNL3*	0	0		0
chr14	105634397	C	-	*JAG2*	0	0		0
chr12	9086909	T	-	*PHC1*	0	0		0
chr4	5758035	C	-	*EVC*	0	0		0
chr1	202396297	T	-	*PPP1R12B*	0	0		0
chr1	202464500	CA	-	*PPP1R12B*	0	0		0
chr15	90294049	G	-	*MESP1*	0	0		0
chr3	14180757	C	-	*TMEM43*	0	0		0
chr5	153857387	G	-	*HAND1*	0	0		0
chr15	90294203	GC	-	*MESP1*	0	0		0
chrX	136649076	C	-	*ZIC3*	0	0		0
chr20	60898582	G	-	*LAMA5*	0	0		0
chr2	220290416	C	-	*DES*	0	0		0
chr12	51589833	C	-	*POU6F1*	0	0		0
chr12	114804154	T	-	*TBX5*	0	0		0
chr4	174450159	C	-	*HAND2*	0	0		0
chr1	151144790	AT	-	*TMOD4*	0	0		0
chr7	73470659	-	T	*ELN*	0	0		0
chr10	94449776	-	G	*HHEX*	0	0		0
chr17	17697130	-	A	*RAI1*	0	0		0
chr16	46744687	-	A	*MYLK3*	0	0		0
chr18	77171468	-	C	*NFATC1*	0	0		0
chr4	114275546	-	C	*ANK2*	0	0		0
chr4	123748299	-	C	*FGF2*	0	0		0
chr12	124820063	-	CAAC	*NCOR2*	0	0		0
chr17	21215536	-	A	*MAP2K3*	0	0		0
chr19	40317590	-	CCCC	*DYRK1B*	0	0		0
chr12	5603608	-	C	*NTF3*	0	0		0
chr12	33031069	-	G	*PKP2*	0	0		0
chr12	52308249	-	G	*ACVRL1*	0	0		0
chr12	115109685	-	A	*TBX3*	0	0		0
chr12	124829420	-	C	*NCOR2*	0	0		0

Table 5.2 Continued

chr11	2869086	-	G	*KCNQ1*	0	0		0
chr10	75758098	-	C	*VCL*	0	0		0
chr17	7130528	-	C	*DVL2*	0	0		0
chr16	3823776	-	G	*CREBBP*	0	0		0
chr16	3828061	-	GG	*CREBBP*	0	0		0
chr15	68612614	-	C	*ITGA11*	0	0		0
chr15	90293392	-	C	*MESP1*	0	0		0
chr19	7119593	-	C	*INSR;INSR*	0	0		0
chr19	40321175	-	C	*DYRK1B*	0	0		0
chr18	19751314	-	G	*GATA6*	0	0		0
chr18	29125917	-	AC	*DSG2*	0	0		0
chr22	31492745	-	G	*SMTN*	0	0		0
chr22	50968118	-	G	*TYMP*	0	0		0
chr20	6750839	-	G	*BMP2*	0	0		0
chr20	10626013	-	G	*JAG1*	0	0		0
chr7	128483879	-	GG	*FLNC*	0	0		0
chr7	128497213	-	C	*FLNC*	0	0		0
chr7	143078671	-	GCCCC	*ZYX*	0	0		0
chr6	50807921	-	TC	*TFAP2B*	0	0		0
chr6	126073017	-	G	*HEY2*	0	0		0
chr5	1880895	-	G	*IRX4*	0	0		0
chr5	155771583	-	GG	*SGCD*	0	0		0
chr4	5800385	-	G	*EVC*	0	0		0
chr4	114274870	-	A	*ANK2*	0	0		0
chr4	114275546	-	CC	*ANK2*	0	0		0
chr3	55508547	-	G	*WNT5A*	0	0		0
chr3	152163311	-	C	*MBNL1*	0	0		0
chr3	193855814	-	G	*HES1*	0	0		0
chr2	121746627	-	GC	*GLI2*	0	0		0
chr1	1273756	-	C	*DVL1*	0	0		0
chr1	16343694	-	G	*HSPB7*	0	0		0
chr1	231557037	-	GC	*EGLN1*	0	0		0
chr1	237777981	-	AA	*RYR2*	0	0		0
chr9	130588044	-	G	*ENG*	0	0		0
chr9	139391543	-	G	*NOTCH1*	0	0		0
chr8	11565886	-	GC	*GATA4*	0	0		0
chr8	80677838	-	C	*HEY1*	0	0		0
chr1	59248148	-	G	*JUN*	0	0		0
chr12	131360223	-	GTAA	*RAN*	0	0		0
chr12	124824739	-	GCCG	*NCOR2*	0	0		0
chr17	29686010	-	G	*NF1*	0	0		0
chr7	128483890	-	C	*FLNC*	0	0		0
chr9	139413208	-	G	*NOTCH1*	0	0		0
chr17	17697121	-	A	*RAI1*	0	0		0
chr6	139694497	-	C	*CITED2*	0	0		0

Table 5.2 Continued

chr17	37883553	-	G	*ERBB2*	0	0		0
chr5	172662040	-	TTTG	*NKX2-5*	0	0		0
chr5	172659759	-	C	*NKX2-5*	0	0		0
chr17	21204186	-	T	*MAP2K3*	0	0		0
chr5	153857391	-	T	*HAND1*	0	0		0
chr12	5154214	-	GG	*KCNA5*	0	0		0
chr12	48185398	-	GG	*HDAC7*	0	0		0
chr11	2869088	-	GC	*KCNQ1*	0	0		0
chr17	39925377	-	G	*JUP*	0	0		0
chr16	3778897	-	C	*CREBBP*	0	0		0
chr14	24845989	-	G	*NFATC4*	0	0		0
chr18	77171472	-	CG	*NFATC1*	0	0		0
chr22	17662874	-	A	*CECR1*	0	0		0
chr20	60884404	-	C	*LAMA5*	0	0		0
chr7	73535521	-	C	*LIMK1*	0	0		0
chr7	128483879	-	G	*FLNC*	0	0		0
chr4	995530	-	G	*IDUA*	0	0		0
chr4	4864617	-	C	*MSX1*	0	0		0
chr4	5620311	-	C	*EVC2*	0	0		0
chr4	5800385	-	GG	*EVC*	0	0		0
chr3	152163311	-	CC	*MBNL1*	0	0		0
chr3	181430234	-	G	*SOX2*	0	0		0
chr2	88387383	-	G	*SMYD1*	0	0		0
chr2	121747180	-	G	*GLI2*	0	0		0
chr2	220283241	-	G	*DES*	0	0		0
chr1	156106155	-	G	*LMNA*	0	0		0
chr1	202318126	-	G	*PPP1R12B*	0	0		0
chr1	208202181	-	C	*PLXNA2*	0	0		0
chr1	226074643	-	GC	*LEFTY1*	0	0		0
chr1	226125353	-	GG	*LEFTY2*	0	0		0
chr1	231557037	-	C	*EGLN1*	0	0		0
chr9	139396883	-	C	*NOTCH1*	0	0		0
chr8	2091344	-	G	*MYOM2*	0	0		0
chr17	7189170	-	G	*SLC2A4*	0	0		0
chr14	73664749	-	GG	*PSEN1*	0	0		0
chr7	4722241	-	C	*FOXK1*	0	0		0
chr7	35288308	-	G	*TBX20*	0	0		0
chr6	1610775	-	G	*FOXC1*	0	0		0
chr6	1612158	-	G	*FOXC1*	0	0		0
chr4	123748299	-	CC	*FGF2*	0	0		0
chr3	193855820	-	G	*HES1*	0	0		0
chr12	98909901	-	G	*TMPO*	0	0		0
chr12	124826453	-	G	*NCOR2*	0	0		0
chr15	23932221	-	G	*NDN*	0	0		0
chr15	48712976	-	G	*FBN1*	0	0		0

Table 5.2 Continued

chr15	48720570	-	G	*FBN1*	0	0		0
chr15	68480091	-	T	*PIAS1*	0	0		0
chr14	105617337	-	G	*JAG2*	0	0		0
chr22	19463076	-	G	*UFD1L*	0	0		0
chr20	6750841	-	GC	*BMP2*	0	0		0
chr7	128478720	-	G	*FLNC*	0	0		0
chr4	5586474	-	C	*EVC2*	0	0		0
chr2	223086008	-	G	*PAX3*	0	0		0
chr1	231557219	-	C	*EGLN1*	0	0		0
chr12	124885147	-	G	*NCOR2*	0	0		0
chr10	88478558	-	C	*LDB3*	0	0		0
chr22	39826174	-	TC	*TAB1*	0	0		0
chr20	60884469	-	C	*LAMA5*	0	0		0
chr20	60899259	-	G	*LAMA5*	0	0		0
chr5	88024395	-	G	*MEF2C*	0	0		0
chr4	114158192	-	G	*ANK2*	0	0		0
chr1	1275458	-	C	*DVL1*	0	0		0
chr17	29686003	-	T	*NF1*	0	0		0
chr2	145156625	-	T	*ZEB2*	0	0		0
chr12	2975681	-	G	*FOXM1*	0	0		0
chr16	3823776	-	GG	*CREBBP*	0	0		0
chr16	51175706	-	G	*SALL1*	0	0		0
chr18	77171472	-	A	*NFATC1*	0	0		0
chr20	60897798	-	G	*LAMA5*	0	0		0
chr1	120548001	-	T	*NOTCH2*	0	0		0
chr10	115804489	-	G	*ADRB1*	0	0		0
chr1	2161015	-	C	*SKI*	0	0		0
chr12	48185398	-	G	*HDAC7*	0	0		0
chr21	48081813	-	C	*PRMT2*	0	0		0
chr3	14526404	-	C	*SLC6A6*	0	0		0
chr1	92185675	-	C	*TGFBR3*	0	0		0
chr8	118830716	-	G	*EXT1*	0	0		0
chr5	1881946	-	G	*IRX4*	0	0		0
chr17	29686007	-	G	*NF1*	0	0		0
chr4	5800389	-	C	*EVC*	0	0		0
chr9	139413094	-	G	*NOTCH1*	0	0		0
chr19	47259732	-	CC	*FKRP*	0	0		0
chr10	101295306	-	G	*NKX2-3*	0	0		0
chr10	115805199	-	G	*ADRB1*	0	0		0
chr17	59482922	-	C	*TBX2*	0	0		0
chr18	19752017	-	G	*GATA6*	0	0		0
chr18	3215058	-	G	*MYOM1*	0	0		0
chr10	101295394	-	G	*NKX2-3*	0	0		0
chr9	141014796	-	C	*CACNA1B*	0	0	rs34080813	0
chr16	54967277	-	G	*IRX5*	0	0		0

Table 5.2 Continued

chr12	133425243	-	C	*CHFR*	0	0		0
chr17	17701501	-	C	*RAI1*	0	0		0
chr16	46744687	-	AA	*MYLK3*	0	0		0
chr2	66739378	-	A	*MEIS1*	0	0		0
chr6	108985582	-	TC	*FOXO3*	0	0		0
chr10	92679010	-	T	*ANKRD1*	0	0		0
chr13	23898659	-	C	*SGCG*	0	0		0
chr12	56092638	-	C	*ITGA7*	0	0		0
chr12	114793568	-	A	*TBX5*	0	0		0
chr10	11356106	-	G	*CELF2*	0	0		0
chr10	75857054	-	G	*VCL*	0	0		0
chr18	3116449	-	G	*MYOM1*	0	0		0
chr4	114275830	-	C	*ANK2*	0	0		0
chr3	152174125	-	C	*MBNL1*	0	0		0
chr2	66739381	-	T	*MEIS1*	0	0		0
chr1	120459193	-	CG	*NOTCH2*	0	0		0
chr17	39925395	-	C	*JUP*	0	0		0
chr17	78449427	-	A	*NPTX1*	0	0		0
chr19	4117556	-	C	*MAP2K2*	0	0		0
chr19	7119592	-	C	*INSR*	0	0		0
chr6	1612231	-	T	*FOXC1*	0	0		0
chr4	995896	-	C	*IDUA*	0	0		0
chr2	40655623	-	T	*SLC8A1*	0	0		0
chr2	145157529	-	C	*ZEB2*	0	0		0
chr9	96715377	-	C	*BARX1*	0	0		0
chr12	52309205	-	C	*ACVRL1*	0	0		0
chr11	2869086	-	GG	*KCNQ1*	0	0		0
chr11	129306804	-	T	*BARX2*	0	0		0
chr18	19751197	-	G	*GATA6*	0	0		0
chr7	44105053	-	C	*PGAM2*	0	0		0
chr7	143085428	-	A	*ZYX*	0	0		0
chr6	108985155	-	G	*FOXO3*	0	0		0
chr1	1991007	-	C	*PRKCZ*	0	0		0
chr9	139413210	-	C	*NOTCH1*	0	0		0
chr12	56092638	-	GC	*ITGA7*	0	0		0
chr10	88659814	-	G	*BMPR1A*	0	0		0
chr17	7189172	-	GT	*SLC2A4*	0	0		0
chr17	29686003	-	TT	*NF1*	0	0		0
chr17	37883560	-	GC	*ERBB2*	0	0		0
chr15	68599984	-	C	*ITGA11*	0	0		0
chr5	122435478	-	G	*PRDM6*	0	0		0
chr5	172659762	-	C	*NKX2-5*	0	0		0
chr4	996233	-	AA	*IDUA*	0	0		0
chr3	193855818	-	A	*HES1*	0	0		0
chr2	239988508	-	C	*HDAC4*	0	0		0

Table 5.2 Continued

chr1	226075310	-	C	*LEFTY1*	0	0			0
chr1	2235396	-	G	*SKI*	0	0			0
chr20	10626015	-	GGTTT	*JAG1*	0	0			0
chr2	239975199	-	G	*HDAC4*	0	0			0
chr7	151372701	-	T	*PRKAG2*	0	0			0
chr12	56088713	-	G	*ITGA7*	0	0			0
chr2	234170	-	C	*SH3YL1*	0	0			0
chr14	105616985	-	G	*JAG2*	0	0			0
chr21	48069627	-	G	*PRMT2*	0	0			0
chr1	2106679	-	C	*PRKCZ*	0	0			0
chr20	10626015	-	GGGTT	*JAG1*	0	0			0
chr2	121747069	-	A	*GLI2*	0	0			0
chr5	139722381	-	T	*HBEGF*	0	0			0
chr9	140772466	-	G	*CACNA1B*	0	0			0
chr18	77211049	-	G	*NFATC1*	0	0			0
chr6	1612155	-	G	*FOXC1*	0	0			0
chr3	71090482	-	C	*FOXP1*	0	0			0
chr1	120464352	-	G	*NOTCH2*	0	0			0
chr15	25616578	-	A	*UBE3A*	0	0			0
chr2	121747180	-	GG	*GLI2*	0	0			0
chr9	140946627	-	G	*CACNA1B*	0	0			0
chr8	119122640	-	C	*EXT1*	0	0			0
chr10	75854057	-	C	*VCL*	0	0			0
chr3	138665524	-	C	*FOXL2*	0	0			0
chr1	237777981	-	TA	*RYR2*	0	0			0
chr10	123263420	-	A	*FGFR2*	0	0			0
chr19	7119590	-	G	*INSR*	0	0			0
chr4	114213583	-	C	*ANK2*	0	0			0
chr14	105617343	-	G	*JAG2*	0	0			0
chr20	10626013	-	GG	*JAG1*	0	0			0
chr9	140946548	-	C	*CACNA1B*	0	0			0
chr6	121768197	-	ATCT	*GJA1*	0	0			0
chr6	139694488	-	C	*CITED2*	0	0			0
chr1	228238594	-	C	*WNT3A*	0	0			0
chr1	59248275	-	GCCC	*JUN*	0	0			0
chr16	3860722	-	G	*CREBBP*	0	0			0
chr3	55508479	-	T	*WNT5A*	0	0			0
chr17	57768006	T	-	*CLTC*	0	0			0
chr7	128481026	T	-	*FLNC*	0	0			0
chr1	156108278	G	-	*LMNA*	0	0			0

Table 5.2 Continued

chr11	1860757	G	-	*TNNI2*	0	0		0
chr17	66508691	-	G	*PRKAR1A*	0	0		0
chr20	33334734	-	AAA	*NCOA6*	0	0		0
chr1	71418662	G	-	*PTGER3*	0	0.0041		0
chr7	35288303	-	G	*TBX20*	0	0		0.002358491
chr2	211179765	-	T	*MYL1*	0	0		0.002358491
chr18	29125917	-	C	*DSG2*	0	0		0.002358491
chr20	33334734	-	A	*NCOA6*	0	0		0.002358491
chr17	21207781	-	T	*MAP2K3*	0	0		0.004716981

Table 5.2. Filtered indels identified using the BWA-Dindel pipeline for the 133 HeartRepair samples.

Chrom.	Position	Ref.	Variant	Gene	EVS5400	1000 Genomes	dbsnp135	control list
chr1	120612003	GG	-	*NOTCH2*	0	0		0
chr1	201334355	T	-	*TNNT2*	0	0		0
chr1	202407190	T	-	*PPP1R12B*	0	0		0
chr10	99338053	G	-	*ANKRD2*	0	0		0
chr2	211179766	T	-	*MYL1*	0	0		0
chr11	112832341	CA	-	*NCAM1*	0	0		0
chr9	139405664	C	-	*NOTCH1*	0	0		0
chr12	124824739	-	GCCG	*NCOR2*	0	0		0
chr3	71090482	-	C	*FOXP1*	0	0		0
chr5	139722381	-	TT	*HBEGF*	0	0		0
chr12	124824739	-	G	*NCOR2*	0	0		0
chr10	88478558	-	C	*LDB3*	0	0		0
chr10	88478560	-	C	*LDB3*	0	0		0
chr12	2983338	-	C	*FOXM1*	0	0		0
chr1	92185675	-	C	*TGFBR3*	0	0		0
chr1	226127214	-	G	*LEFTY2*	0	0		0
chr2	121747069	-	A	*GLI2*	0	0		0
chr6	50807921	-	C	*TFAP2B*	0	0		0
chr11	2869086	-	GG	*KCNQ1*	0	0		0
chr11	2869088	-	GC	*KCNQ1*	0	0		0
chr12	115109685	-	A	*TBX3*	0	0		0
chr14	73664749	-	GG	*PSEN1*	0	0		0
chr16	3778897	-	C	*CREBBP*	0	0		0
chr20	6750839	-	G	*BMP2*	0	0		0
chr4	123748299	-	C	*FGF2*	0	0		0
chr12	124885147	-	G	*NCOR2*	0	0		0
chr10	92679010	-	T	*ANKRD1*	0	0		0
chr2	66739381	-	T	*MEIS1*	0	0		0
chr1	2235396	-	G	*SKI*	0	0		0
chr9	140773612	-	A	*CACNA1B*	0	0		0
chr9	140777194	A	-	*CACNA1B*	0	0		0
chr1	71418662	G	-	*PTGER3*	0	0.0041		0
chr4	168155291	CA	-	*SPOCK3*	0	0.01		0
chr2	211179765	-	T	*MYL1*	0	0		0.002358491
chr20	33334734	-	A	*NCOA6*	0	0		0.002358491

Table 5.3. Filtered indels identified using the BWA-GATK pipeline for the 133 HeartRepair samples.

References

Acharya, A., Hans, C.P., Koenig, S.N., Nichols, H.A., Galindo, C.L., Garner, H.R., Merrill, W.H., Hinton, R.B. and Garg, V. (2011) 'Inhibitory role of Notch1 in calcific aortic valve disease', *PLoS One*, 6(11), p. e27743.

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 Methods*, 7(4), pp. 248-9.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H. and Durbin, R. (2011) 'Dindel: accurate indel calls from short-read data', *Genome Res*, 21(6), pp. 961-73.

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', *Eur J Hum Genet*, 10(1), pp. 17-25.

Arking, D.E. and Chakravarti, A. (2009) 'Understanding cardiovascular disease through the lens of genome-wide association studies', *Trends Genet*, 25(9), pp. 387-94.

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., Yang, H. and Zhang, X. (2011) 'Comprehensive comparison of three commercial human whole-exome capture platforms', *Genome Biol*, 12(9), p. R95.

Ashburn, D.A., Blackstone, E.H., Wells, W.J., Jonas, R.A., Pigula, F.A., Manning, P.B., Lofland, G.K., Williams, W.G., McCrindle, B.W. and Congenital Heart Surgeons Study, m. (2004) 'Determinants of mortality and type of repair in neonates with pulmonary atresia and intact ventricular septum', *J Thorac Cardiovasc Surg*, 127(4), pp. 1000-7; discussion 1007-8.

Attenhofer Jost, C.H., Connolly, H.M., Dearani, J.A., Edwards, W.D. and Danielson, G.K. (2007) 'Ebstein's anomaly', *Circulation*, 115(2), pp. 277-85.

Audo, I., Bujakowska, K.M., Leveillard, T., Mohand-Said, S., Lancelot, M.E., Germain, A., Antonio, A., Michiels, C., Saraiva, J.P., Letexier, M., Sahel, J.A., Bhattacharya, S.S. and Zeitz, C. (2012) 'Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases', *Orphanet J Rare Dis*, 7, p. 8.

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

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. 745-55.

Bansal, V. and Libiger, O. (2011) 'A probabilistic method for the detection and genotyping of small indels from population-scale sequence data', *Bioinformatics*, 27(15), pp. 2047-53.

Barton, P.J., Cohen, A., Robert, B., Fiszman, M.Y., Bonhomme, F., Guenet, J.L., Leader, D.P. and Buckingham, M.E. (1985) 'The myosin alkali light chains of mouse ventricular and slow skeletal muscle are indistinguishable and are encoded by the same gene', *J Biol Chem*, 260(14), pp. 8578-84.

Basson, C.T., Cowley, G.S., Solomon, S.D., Weissman, B., Poznanski, A.K., Traill, T.A., Seidman, J.G. and Seidman, C.E. (1994) 'The clinical and genetic spectrum of the Holt-Oram syndrome (heart-hand syndrome)', *N Engl J Med*, 330(13), pp. 885-91.

Becerra, J.E., Khoury, M.J., Cordero, J.F. and Erickson, J.D. (1990) 'Diabetes mellitus during pregnancy and the risks for specific birth defects: a population-based case-control study', *Pediatrics*, 85(1), pp. 1-9.

Beffagna, G., De Bortoli, M., Nava, A., Salamon, M., Lorenzon, A., Zaccolo, M., Mancuso, L., Sigalotti, L., Bauce, B., Occhi, G., Basso, C., Lanfranchi, G., Towbin, J.A., Thiene, G., Danieli, G.A. and Rampazzo, A. (2007) 'Missense mutations in desmocollin-2 N-terminus, associated with arrhythmogenic right ventricular cardiomyopathy, affect intracellular localization of desmocollin-2 in vitro', *BMC Med Genet*, 8, p. 65.

Belgrano, A., Rakicevic, L., Mittempergher, L., Campanaro, S., Martinelli, V.C., Mouly, V., Valle, G., Kojic, S. and Faulkner, G. (2011) 'Multi-tasking role of the mechanosensing protein Ankrd2 in the signaling network of striated muscle', *PLoS One*, 6(10), p. e25519.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., et al. (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218), pp. 53-9.
Bhangale, T.R., Rieder, M.J., Livingston, R.J. and Nickerson, D.A. (2005) 'Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes', *Hum Mol Genet*, 14(1), pp. 59-69.

Bolze, A., Byun, M., McDonald, D., Morgan, N.V., Abhyankar, A., Premkumar, L., Puel, A., Bacon, C.M., Rieux-Laucat, F., Pang, K., Britland, A., Abel, L., Cant, A., Maher, E.R., Riedl, S.J., Hambleton, S. and Casanova, J.L. (2010) 'Whole-exome-sequencing-based discovery of human FADD deficiency', *Am J Hum Genet*, 87(6), pp. 873-81.

Boneva, R.S., Botto, L.D., Moore, C.A., Yang, Q., Correa, A. and Erickson, J.D. (2001) 'Mortality associated with congenital heart defects in the United States: trends and racial disparities, 1979-1997', *Circulation*, 103(19), pp. 2376-81.

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

Bull, C., de Leval, M.R., Mercanti, C., Macartney, F.J. and Anderson, R.H. (1982) 'Pulmonary atresia and intact ventricular septum: a revised classification', *Circulation*, 66(2), pp. 266-72.

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', *Lancet*, 351(9099), pp. 311-6.

Carmignac, V., Thevenon, J., Ades, L., Callewaert, B., Julia, S., Thauvin-Robinet, C., Gueneau, L., Courcet, J.B., Lopez, E., Holman, K., Renard, M., Plauchu, H., Plessis, G., De Backer, J., Child, A., Arno, G., Duplomb, L., Callier, P., Aral, B., Vabres, P., Gigot, N., Arbustini, E., Grasso, M., Robinson, P.N., Goizet, C., Baumann, C., Di Rocco, M., Sanchez Del Pozo, J., Huet, F., Jondeau, G., Collod-Beroud, G., Beroud, C., Amiel, J., Cormier-Daire, V., Riviere, J.B., Boileau, C., De Paepe, A. and Faivre, L. (2012) 'In-Frame Mutations in Exon 1 of SKI Cause Dominant Shprintzen-Goldberg Syndrome', *Am J Hum Genet*, 91(5), pp. 950-7.

Cartwright, R.A. (2009) 'Problems and solutions for estimating indel rates and length distributions', *Mol Biol Evol*, 26(2), pp. 473-80.

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.

Chan, P.A., Duraisamy, S., Miller, P.J., Newell, J.A., McBride, C., Bond, J.P., Raevaara, T., Ollila, S., Nystrom, M., Grimm, A.J., Christodoulou, J., Oetting, W.S. and Greenblatt, M.S. (2007) 'Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR)', *Hum Mutat*, 28(7), pp. 683-93.
Chen, J.M., Ferec, C. and Cooper, D.N. (2010) 'Revealing the human mutome', *Clin Genet*, 78(4), pp. 310-20.

Ching, Y.H., Ghosh, T.K., Cross, S.J., Packham, E.A., Honeyman, L., Loughna, S., Robinson, T.E., Dearlove, A.M., Ribas, G., Bonser, A.J., Thomas, N.R., Scotter, A.J., Caves, L.S., Tyrrell, G.P., Newbury-Ecob, R.A., Munnich, A., Bonnet, D. and Brook, J.D. (2005) 'Mutation in myosin heavy chain 6 causes atrial septal defect', *Nat Genet*, 37(4), pp. 423-8.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. and Lifton, R.P. (2009) 'Genetic diagnosis by whole exome capture and massively parallel DNA sequencing', *Proc Natl Acad Sci U S A*, 106(45), pp. 19096-101.

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. 415-25.

Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Euskirchen, G., Butte, A.J. and Snyder, M. (2011) 'Performance comparison of exome DNA sequencing technologies', *Nat Biotechnol*, 29(10), pp. 908-14.

Codd, M.B., Sugrue, D.D., Gersh, B.J. and Melton, L.J., 3rd (1989) 'Epidemiology of idiopathic dilated and hypertrophic cardiomyopathy. A population-based study in Olmsted County, Minnesota, 1975-1984', *Circulation*, 80(3), pp. 564-72.

Coonrod, E.M., Durtschi, J.D., Margraf, R.L. and Voelkerding, K.V. (2012) 'Developing Genome and Exome Sequencing for Candidate Gene Identification in Inherited Disorders', *Arch Pathol Lab Med*.

Cooper, D.N., Chen, J.M., Ball, E.V., Howells, K., Mort, M., Phillips, A.D., Chuzhanova, N., Krawczak, M., Kehrer-Sawatzki, H. and Stenson, P.D. (2010) 'Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics', *Hum Mutat*, 31(6), pp. 631-55.

Cooper, W.O., Hernandez-Diaz, S., Arbogast, P.G., Dudley, J.A., Dyer, S., Gideon, P.S., Hall, K. and Ray, W.A. (2006) 'Major congenital malformations after first-trimester exposure to ACE inhibitors', *N Engl J Med*, 354(23), pp. 2443-51.

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

Correa-Villasenor, A., Ferencz, C., Neill, C.A., Wilson, P.D. and Boughman, J.A. (1994) 'Ebstein's malformation of the tricuspid valve: genetic and environmental factors. The Baltimore-Washington Infant Study Group', *Teratology*, 50(2), pp. 137-47.

Craig, B. (2006) 'Atrioventricular septal defect: from fetus to adult', *Heart*, 92(12), pp. 1879-85.
Cresci, M., Vecoli, C., Foffa, I., Pulignani, S., Ait-Ali, L. and Andreassi, M.G. (2012) 'Lack of Association of the 3'-UTR Polymorphism (rs1017) in the ISL1 Gene and Risk of Congenital Heart Disease in the White Population', *Pediatr Cardiol*.

Day-Williams, A.G. and Zeggini, E. (2011) 'The effect of next-generation sequencing technology on complex trait research', *Eur J Clin Invest*, 41(5), pp. 561-7.

De Bortoli, M., Beffagna, G., Bauce, B., Lorenzon, A., Smaniotto, G., Rigato, I., Calore, M., Li Mura, I.E., Basso, C., Thiene, G., Lanfranchi, G., Danieli, G.A., Nava, A. and Rampazzo, A. (2010) 'The p.A897KfsX4 frameshift variation in desmocollin-2 is not a causative mutation in arrhythmogenic right ventricular cardiomyopathy', *Eur J Hum Genet*, 18(7), pp. 776-82.

De Keulenaer, S., Hellemans, J., Lefever, S., Renard, J.P., De Schrijver, J., Van de Voorde, H., Tabatabaiefar, M.A., Van Nieuwerburgh, F., Flamez, D., Pattyn, F., Scharlaken, B., Deforce, D., Bekaert, S., Van Criekinge, W., Vandesompele, J., Van Camp, G. and Coucke, P. (2012) 'Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform', *BMC Med Genomics*, 5, p. 17.

de la Pompa, J.L. and Epstein, J.A. (2012) 'Coordinating tissue interactions: Notch signaling in cardiac development and disease', *Dev Cell*, 22(2), pp. 244-54.

Dellefave, L. and McNally, E.M. (2010) 'The genetics of dilated cardiomyopathy', *Curr Opin Cardiol*, 25(3), pp. 198-204.

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', *Nat Genet*, 43(5), pp. 491-8.

Dib, C., Araoz, P.A., Davies, N.P., Dearani, J.A. and Ammash, N.M. (2012) 'Hypoplastic right-heart syndrome presenting as multiple miscarriages', *Tex Heart Inst J*, 39(2), pp. 249-54.

Dickinson, R.E., Griffin, H., Bigley, V., Reynard, L.N., Hussain, R., Haniffa, M., Lakey, J.H., Rahman, T., Wang, X.N., McGovern, N., Pagan, S., Cookson, S., McDonald, D., Chua, I., Wallis, J., Cant, A., Wright, M., Keavney, B., Chinnery, P.F., Loughlin, J., Hambleton, S., Santibanez-Koref, M. and Collin, M. (2011) 'Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency', *Blood*, 118(10), pp. 2656-8.

Digilio, M.C., Bernardini, L., Lepri, F., Giuffrida, M.G., Guida, V., Baban, A., Versacci, P., Capolino, R., Torres, B., De Luca, A., Novelli, A., Marino, B. and Dallapiccola, B. (2011) 'Ebstein anomaly: Genetic heterogeneity and association with microdeletions 1p36 and 8p23.1', *Am J Med Genet A*, 155A(9), pp. 2196-202.

Digilio, M.C., Marino, B., Toscano, A., Giannotti, A. and Dallapiccola, B. (1999) 'Atrioventricular canal defect without Down syndrome: a heterogeneous malformation', *Am J Med Genet*, 85(2), pp. 140-6.

Drielsma, A., Jalas, C., Simonis, N., Desir, J., Simanovsky, N., Pirson, I., Elpeleg, O., Abramowicz, M. and Edvardson, S. (2012) 'Two novel CCDC88C mutations confirm the role of DAPLE in autosomal recessive congenital hydrocephalus', *J Med Genet*, 49(11), pp. 708-12.

Duarte, A., Hirashima, M., Benedito, R., Trindade, A., Diniz, P., Bekman, E., Costa, L., Henrique, D. and Rossant, J. (2004) 'Dosage-sensitive requirement for mouse Dll4 in artery development', *Genes Dev*, 18(20), pp. 2474-8.

Dymecki, S.M., Niederhuber, J.E. and Desiderio, S.V. (1990) 'Specific expression of a tyrosine kinase gene, blk, in B lymphoid cells', *Science*, 247(4940), pp. 332-6.

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', *Hum Mol Genet*, 10(2), pp. 163-9.

Engberding, R., Stollberger, C., Ong, P., Yelbuz, T.M., Gerecke, B.J. and Breithardt, G. (2010) 'Isolated non-compaction cardiomyopathy', *Dtsch Arztebl Int*, 107(12), pp. 206-13.
England, J. and Loughna, S. (2012) 'Heavy and light roles: myosin in the morphogenesis of the heart', *Cell Mol Life Sci*.

Erlich, Y., Edvardson, S., Hodges, E., Zenvirt, S., Thekkat, P., Shaag, A., Dor, T., Hannon, G.J. and Elpeleg, O. (2011) 'Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis', *Genome Res*, 21(5), pp. 658-64.
Faita, F., Vecoli, C., Foffa, I. and Andreassi, M.G. (2012) 'Next generation sequencing in cardiovascular diseases', *World J Cardiol*, 4(10), pp. 288-95.

Feng, T. and Zhu, X. (2010) 'Genome-wide searching of rare genetic variants in WTCCC data', *Hum Genet*, 128(3), pp. 269-80.

Ferencz, C., Neill, C.A., Boughman, J.A., Rubin, J.D., Brenner, J.I. and Perry, L.W. (1989) 'Congenital cardiovascular malformations associated with chromosome abnormalities: an epidemiologic study', *J Pediatr*, 114(1), pp. 79-86.

Forissier, J.F., Carrier, L., Farza, H., Bonne, G., Bercovici, J., Richard, P., Hainque, B., Townsend, P.J., Yacoub, M.H., Faure, S., Dubourg, O., Millaire, A., Hagege, A.A., Desnos, M., Komajda, M.

and Schwartz, K. (1996) 'Codon 102 of the cardiac troponin T gene is a putative hot spot for mutations in familial hypertrophic cardiomyopathy', *Circulation*, 94(12), pp. 3069-73.
Fruitman, D.S. (2000) 'Hypoplastic left heart syndrome: Prognosis and management options', *Paediatr Child Health*, 5(4), pp. 219-25.

Fuchs-Telem, D., Sarig, O., van Steensel, M.A., Isakov, O., Israeli, S., Nousbeck, J., Richard, K., Winnepenninckx, V., Vernooij, M., Shomron, N., Uitto, J., Fleckman, P., Richard, G. and Sprecher, E. (2012) 'Familial pityriasis rubra pilaris is caused by mutations in CARD14', *Am J Hum Genet*, 91(1), pp. 163-70.

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. 270-4.

Gehmlich, K., Syrris, P., Peskett, E., Evans, A., Ehler, E., Asimaki, A., Anastasakis, A., Tsatsopoulou, A., Vouliotis, A.I., Stefanadis, C., Saffitz, J.E., Protonotarios, N. and McKenna, W.J. (2011) 'Mechanistic insights into arrhythmogenic right ventricular cardiomyopathy caused by desmocollin-2 mutations', *Cardiovasc Res*, 90(1), pp. 77-87.

Gilissen, C., Hoischen, A., Brunner, H.G. and Veltman, J.A. (2011) 'Unlocking Mendelian disease using exome sequencing', *Genome Biol*, 12(9), p. 228.

Goh, K., Sasajima, T., Inaba, M., Yamamoto, H., Kawashima, E. and Kubo, Y. (1998) 'Isolated right ventricular hypoplasia: intraoperative balloon occlusion test', *Ann Thorac Surg*, 65(2), pp. 551-3.

Gordon, W.R., Arnett, K.L. and Blacklow, S.C. (2008) 'The molecular logic of Notch signaling--a structural and biochemical perspective', *J Cell Sci*, 121(Pt 19), pp. 3109-19.

Grau Salvat, C., Pont, V., Cors, J.R. and Aliaga, A. (1999) 'Hereditary sclerosing poikiloderma of Weary: report of a new case', *Br J Dermatol*, 140(2), pp. 366-8.

Green, E.K., Priestley, M.D., Waters, J., Maliszewska, C., Latif, F. and Maher, E.R. (2000) 'Detailed mapping of a congenital heart disease gene in chromosome 3p25', *J Med Genet*, 37(8), pp. 581-7.

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 Mde, L., 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', *Nat Genet*, 41(8), pp. 931-5.

Griffin, H.R., Hall, D.H., Topf, A., Eden, J., Stuart, A.G., Parsons, J., Peart, I., Deanfield, J.E., O'Sullivan, J., Babu-Narayan, S.V., Gatzoulis, M.A., Bu'lock, F.A., Bhattacharya, S., Bentham, J., Farrall, M., Riveron, J.G., Brook, J.D., Burn, J., Cordell, H.J., Goodship, J.A. and Keavney, B. (2009) 'Genetic variation in VEGF does not contribute significantly to the risk of congenital cardiovascular malformation', *PLoS One*, 4(3), p. e4978.

Grossfeld, P. (2007) 'Hypoplastic left heart syndrome: new insights', *Circ Res*, 100(9), pp. 1246-8.

Gutgesell, H.P. (1975) 'Pulmonary Valve Atresia with Intact Ventricular Septum', *Cardiovasc Dis*, 2(2), pp. 148-155.

Hager, A., Kaemmerer, H., Eicken, A., Fratz, S. and Hess, J. (2002) 'Long-term survival of patients with univentricular heart not treated surgically', *J Thorac Cardiovasc Surg*, 123(6), pp. 1214-7.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. and Frazer, K.A. (2009) 'Evaluation of next generation sequencing platforms for population targeted sequencing studies', *Genome Biol*, 10(3), p. R32.
Harland, M., Mistry, S., Bishop, D.T. and Bishop, J.A. (2001) 'A deep intronic mutation in CDKN2A is associated with disease in a subset of melanoma pedigrees', *Hum Mol Genet*, 10(23), pp. 2679-86.

Hartman, R.J., Riehle-Colarusso, T., Lin, A., Frias, J.L., Patel, S.S., Duwe, K., Correa, A. and Rasmussen, S.A. (2011) 'Descriptive study of nonsyndromic atrioventricular septal defects in the National Birth Defects Prevention Study, 1997-2005', *Am J Med Genet A*, 155A(3), pp. 555-64.

Hazebroek, M., Dennert, R. and Heymans, S. (2012) 'Idiopathic dilated cardiomyopathy: possible triggers and treatment strategies', *Neth Heart J*.

Herman, D.S., Lam, L., Taylor, M.R., Wang, L., Teekakirikul, P., Christodoulou, D., Conner, L., DePalma, S.R., McDonough, B., Sparks, E., Teodorescu, D.L., Cirino, A.L., Banner, N.R., Pennell, D.J., Graw, S., Merlo, M., Di Lenarda, A., Sinagra, G., Bos, J.M., Ackerman, M.J., Mitchell, R.N., Murry, C.E., Lakdawala, N.K., Ho, C.Y., Barton, P.J., Cook, S.A., Mestroni, L., Seidman, J.G. and Seidman, C.E. (2012) 'Truncations of titin causing dilated cardiomyopathy', *N Engl J Med*, 366(7), pp. 619-28.

Hershberger, R.E., Morales, A. and Siegfried, J.D. (2010) 'Clinical and genetic issues in dilated cardiomyopathy: a review for genetics professionals', *Genet Med*, 12(11), pp. 655-67.
Hickey, E.J., Caldarone, C.A. and McCrindle, B.W. (2012) 'Left ventricular hypoplasia: a spectrum of disease involving the left ventricular outflow tract, aortic valve, and aorta', *J Am Coll Cardiol*, 59(1 Suppl), pp. S43-54.

Hicks, S., Wheeler, D.A., Plon, S.E. and Kimmel, M. (2011) 'Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed', *Hum Mutat*, 32(6), pp. 661-8.

High, F.A. and Epstein, J.A. (2008) 'The multifaceted role of Notch in cardiac development and disease', *Nat Rev Genet*, 9(1), pp. 49-61.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc Natl Acad Sci U S A*, 106(23), pp. 9362-7.

Hinton, R.B., Martin, L.J., Rame-Gowda, S., Tabangin, M.E., Cripe, L.H. and Benson, D.W. (2009) 'Hypoplastic left heart syndrome links to chromosomes 10q and 6q and is genetically related to bicuspid aortic valve', *J Am Coll Cardiol*, 53(12), pp. 1065-71.

Hirschhorn, J.N. and Daly, M.J. (2005) 'Genome-wide association studies for common diseases and complex traits', *Nat Rev Genet*, 6(2), pp. 95-108.

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

Horvath, R., Holinski-Feder, E., Neeve, V.C., Pyle, A., Griffin, H., Ashok, D., Foley, C., Hudson, G., Rautenstrauss, B., Nurnberg, G., Nurnberg, P., Kortler, J., Neitzel, B., Bassmann, I., Rahman, T., Keavney, B., Loughlin, J., Hambleton, S., Schoser, B., Lochmuller, H., Santibanez-Koref, M. and Chinnery, P.F. (2012) 'A new phenotype of brain iron accumulation with dystonia, optic atrophy, and peripheral neuropathy', *Mov Disord*, 27(6), pp. 789-93.

Hu, J. and Ng, P.C. (2012) 'Predicting the effects of frameshifting indels', *Genome Biol*, 13(2), p. R9.

Huang, C., Sheikh, F., Hollander, M., Cai, C., Becker, D., Chu, P.H., Evans, S. and Chen, J. (2003) 'Embryonic atrial function is essential for mouse embryogenesis, cardiac morphogenesis and angiogenesis', *Development*, 130(24), pp. 6111-9.

Huang, W., Zhang, R. and Xu, X. (2009) 'Myofibrillogenesis in the developing zebrafish heart: A functional study of tnnt2', *Dev Biol*, 331(2), pp. 237-49.

Ichida, F., Tsubata, S., Bowles, K.R., Haneda, N., Uese, K., Miyawaki, T., Dreyer, W.J., Messina, J., Li, H., Bowles, N.E. and Towbin, J.A. (2001) 'Novel gene mutations in patients with left ventricular noncompaction or Barth syndrome', *Circulation*, 103(9), pp. 1256-63.
Islam, K.B., Rabbani, H., Larsson, C., Sanders, R. and Smith, C.I. (1995) 'Molecular cloning, characterization, and chromosomal localization of a human lymphoid tyrosine kinase related to murine Blk', *J Immunol*, 154(3), pp. 1265-72.

Jamieson, C.R., van der Burgt, I., Brady, A.F., van Reen, M., Elsawi, M.M., Hol, F., Jeffery, S., Patton, M.A. and Mariman, E. (1994) 'Mapping a gene for Noonan syndrome to the long arm of chromosome 12', *Nat Genet*, 8(4), pp. 357-60.

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.

Ji, H.P. (2012) 'Improving bioinformatic pipelines for exome variant calling', *Genome Med*, 4(1), p. 7.

Jobard, F., Lefevre, C., Karaduman, A., Blanchet-Bardon, C., Emre, S., Weissenbach, J., Ozguc, M., Lathrop, M., Prud'homme, J.F. and Fischer, J. (2002) 'Lipoxygenase-3 (ALOXE3) and 12(R)-lipoxygenase (ALOX12B) are mutated in non-bullous congenital ichthyosiform erythroderma (NCIE) linked to chromosome 17p13.1', *Hum Mol Genet*, 11(1), pp. 107-13.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., Ding, J., McCluskey, L., Martinez-Lage, M., Falcone, D., Hernandez, D.G., Arepalli, S., Chong, S., Schymick, J.C., Rothstein, J., Landi, F., Wang, Y.D., Calvo, A., Mora, G., Sabatelli, M., Monsurro, M.R., Battistini, S., Salvi, F., Spataro, R., Sola, P., Borghero, G., Consortium, I., Galassi, G., Scholz, S.W., Taylor, J.P., Restagno, G., Chio, A. and

Traynor, B.J. (2010a) 'Exome sequencing reveals VCP mutations as a cause of familial ALS', *Neuron*, 68(5), pp. 857-64.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., Ding, J., McCluskey, L., Martinez-Lage, M., Falcone, D., Hernandez, D.G., Arepalli, S., Chong, S., Schymick, J.C., Rothstein, J., Landi, F., Wang, Y.D., Calvo, A., Mora, G., Sabatelli, M., Monsurro, M.R., Battistini, S., Salvi, F., Spataro, R., Sola, P., Borghero, G., Galassi, G., Scholz, S.W., Taylor, J.P., Restagno, G., Chio, A. and Traynor, B.J. (2010b) 'Exome sequencing reveals VCP mutations as a cause of familial ALS', *Neuron*, 68(5), pp. 857-64.

Jordan, D.M., Ramensky, V.E. and Sunyaev, S.R. (2010) 'Human allelic variation: perspective from protein function, structure, and evolution', *Curr Opin Struct Biol*, 20(3), pp. 342-50.
Juran, B.D. and Lazaridis, K.N. (2011) 'Genomics in the post-GWAS era', *Semin Liver Dis*, 31(2), pp. 215-22.

Kamisago, M., Sharma, S.D., DePalma, S.R., Solomon, S., Sharma, P., McDonough, B., Smoot, L., Mullen, M.P., Woolf, P.K., Wigle, E.D., Seidman, J.G. and Seidman, C.E. (2000) 'Mutations in sarcomere protein genes as a cause of dilated cardiomyopathy', *N Engl J Med*, 343(23), pp. 1688-96.

Karchin, R. (2009) 'Next generation tools for the annotation of human SNPs', *Brief Bioinform*, 10(1), pp. 35-52.

Khairy, P., Poirier, N. and Mercier, L.A. (2007) 'Univentricular heart', *Circulation*, 115(6), pp. 800-12.

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?', *Br J Dermatol*, 155(5), pp. 1057-61.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., Hultman, C.M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y.Y., Price, A.L., de Bakker, P.I., Purcell, S.M. and Sunyaev, S.R. (2012) 'Exome sequencing and the genetic basis of complex traits', *Nat Genet*, 44(6), pp. 623-30.

Klaassen, S., Probst, S., Oechslin, E., Gerull, B., Krings, G., Schuler, P., Greutmann, M., Hurlimann, D., Yegitbasi, M., Pons, L., Gramlich, M., Drenckhahn, J.D., Heuser, A., Berger, F., Jenni, R. and Thierfelder, L. (2008) 'Mutations in sarcomere protein genes in left ventricular noncompaction', *Circulation*, 117(22), pp. 2893-901.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K. and Ding, L. (2009) 'VarScan: variant detection in massively parallel sequencing of individual and pooled samples', *Bioinformatics*, 25(17), pp. 2283-5.

Koboldt, D.C., Zhang, Q., 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 Res*, 22(3), pp. 568-76.

Krantz, I.D., Smith, R., Colliton, R.P., Tinkel, H., Zackai, E.H., Piccoli, D.A., Goldmuntz, E. and Spinner, N.B. (1999) 'Jagged1 mutations in patients ascertained with isolated congenital heart defects', *Am J Med Genet*, 84(1), pp. 56-60.

Krawitz, P.M., Schweiger, M.R., Rodelsperger, C., Marcelis, C., Kolsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., Isau, M., Fischer, A., Dahl, A., Kerick, M., Hecht, J., Kohler, S., Jager, M., Grunhagen, J., de Condor, B.J., Doelken, S., Brunner, H.G., Meinecke, P., Passarge, E., Thompson, M.D., Cole, D.E., Horn, D., Roscioli, T., Mundlos, S. and Robinson, P.N. (2010) 'Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome', *Nat Genet*, 42(10), pp. 827-9.

Krebs, L.T., Xue, Y., Norton, C.R., Shutter, J.R., Maguire, M., Sundberg, J.P., Gallahan, D., Closson, V., Kitajewski, J., Callahan, R., Smith, G.H., Stark, K.L. and Gridley, T. (2000) 'Notch signaling is essential for vascular morphogenesis in mice', *Genes Dev*, 14(11), pp. 1343-52.
Kryukov, G.V., Pennacchio, L.A. and Sunyaev, S.R. (2007) 'Most rare missense alleles are deleterious in humans: implications for complex disease and association studies', *Am J Hum Genet*, 80(4), pp. 727-39.

Ku, C.S., Naidoo, N. and Pawitan, Y. (2011) 'Revisiting Mendelian disorders through exome sequencing', *Hum Genet*, 129(4), pp. 351-70.

Kumar, A., Williams, C.A. and Victorica, B.E. (1994) 'Familial atrioventricular septal defect: possible genetic mechanisms', *Br Heart J*, 71(1), pp. 79-81.

Kumar, P., Henikoff, S. and Ng, P.C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nat Protoc*, 4(7), pp. 1073-81.

Kyndt, F., Gueffet, J.P., Probst, V., Jaafar, P., Legendre, A., Le Bouffant, F., Toquet, C., Roy, E., McGregor, L., Lynch, S.A., Newbury-Ecob, R., Tran, V., Young, I., Trochu, J.N., Le Marec, H. and Schott, J.J. (2007) 'Mutations in the gene encoding filamin A as a cause for familial cardiac valvular dystrophy', *Circulation*, 115(1), pp. 40-9.

Lakdawala, N.K., Funke, B.H., Baxter, S., Cirino, A.L., Roberts, A.E., Judge, D.P., Johnson, N., Mendelsohn, N.J., Morel, C., Care, M., Chung, W.K., Jones, C., Psychogios, A., Duffy, E., Rehm, H.L., White, E., Seidman, J.G., Seidman, C.E. and Ho, C.Y. (2012) 'Genetic testing for dilated cardiomyopathy in clinical practice', *J Card Fail*, 18(4), pp. 296-303.

Lalonde, E., Albrecht, S., Ha, K.C., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J. and Jabado, N. (2010) 'Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing', *Hum Mutat*, 31(8), pp. 918-23.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860-921.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9(4), pp. 357-9.

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.
Ledergerber, C. and Dessimoz, C. (2011) 'Base-calling for next-generation sequencing platforms', *Brief Bioinform*, 12(5), pp. 489-97.

Lee, H.J., Shin, D.H., Choi, J.S. and Kim, K.H. (2012) 'Hereditary sclerosing poikiloderma', *J Korean Med Sci*, 27(2), pp. 225-7.

Lemos, R.R., Souza, M.B. and Oliveira, J.R. (2012) 'Exploring the Implications of INDELs in Neuropsychiatric Genetics: Challenges and Perspectives', *J Mol Neurosci*.

Levy, H.L., Guldberg, P., Guttler, F., Hanley, W.B., Matalon, R., Rouse, B.M., Trefz, F., Azen, C., Allred, E.N., de la Cruz, F. and Koch, R. (2001) 'Congenital heart disease in maternal phenylketonuria: report from the Maternal PKU Collaborative Study', *Pediatr Res*, 49(5), pp. 636-42.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. and Venter, J.C. (2007) 'The diploid genome sequence of an individual human', *PLoS Biol*, 5(10), p. e254.

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. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078-9.

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

Li, L., Krantz, I.D., Deng, Y., Genin, A., Banta, A.B., Collins, C.C., Qi, M., Trask, B.J., Kuo, W.L., Cochran, J., Costa, T., Pierpont, M.E., Rand, E.B., Piccoli, D.A., Hood, L. and Spinner, N.B. (1997) 'Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1', *Nat Genet*, 16(3), pp. 243-51.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b) 'SOAP: short oligonucleotide alignment program', *Bioinformatics*, 24(5), pp. 713-4.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) 'Comparison of next-generation sequencing systems', *J Biomed Biotechnol*, 2012, p. 251364.
Locke, A.E., Dooley, K.J., Tinker, S.W., Cheong, S.Y., Feingold, E., Allen, E.G., Freeman, S.B., Torfs, C.P., Cua, C.L., Epstein, M.P., Wu, M.C., Lin, X., Capone, G., Sherman, S.L. and Bean, L.J. (2010) 'Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome', *Genet Epidemiol*, 34(6), pp. 613-23.

Luk, A., Ahn, E., Soor, G.S. and Butany, J. (2009) 'Dilated cardiomyopathy: a review', *J Clin Pathol*, 62(3), pp. 219-25.

Lunter, G. and Goodson, M. (2011) 'Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads', *Genome Res*, 21(6), pp. 936-9.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.M., Hunt, T., Barnes, I.H., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B. and Tyler-Smith, C. (2012) 'A systematic survey of loss-of-function variants in human protein-coding genes', *Science*, 335(6070), pp. 823-8.

MacGrogan, D., Luna-Zurita, L. and de la Pompa, J.L. (2011) 'Notch signaling in cardiac valve development and disease', *Birth Defects Res A Clin Mol Teratol*, 91(6), pp. 449-59.
Maher, B. (2008) 'Personal genomes: The case of the missing heritability', *Nature*, 456(7218), pp. 18-21.

Mahon, N.G., Murphy, R.T., MacRae, C.A., Caforio, A.L., Elliott, P.M. and McKenna, W.J. (2005) 'Echocardiographic evaluation in asymptomatic relatives of patients with dilated cardiomyopathy reveals preclinical disease', *Ann Intern Med*, 143(2), pp. 108-15.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B. and de los Campos, G. (2011) 'Beyond missing heritability: prediction of complex traits', *PLoS Genet*, 7(4), p. e1002051.

Malhis, N., Butterfield, Y.S., Ester, M. and Jones, S.J. (2009) 'Slider--maximum use of probability information for alignment of short sequence reads and SNP detection', *Bioinformatics*, 25(1), pp. 6-13.

Manning, N., Kaufman, L. and Roberts, P. (2005) 'Genetics of cardiological disorders', *Semin Fetal Neonatal Med*, 10(3), pp. 259-69.

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. (2008) 'The impact of next-generation sequencing technology on genetics', *Trends Genet*, 24(3), pp. 133-41.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., Fulton, L.A., Locke, D.P., Magrini, V.J., Abbott, R.M., Vickery, T.L., Reed, J.S., Robinson, J.S., Wylie, T., Smith, S.M., Carmichael, L., Eldred, J.M., Harris, C.C., Walker, J., Peck, J.B., Du, F., Dukes, A.F., Sanderson, G.E., Brummett, A.M., Clark, E., McMichael, J.F., Meyer, R.J., Schindler, J.K., Pohl, C.S., Wallis, J.W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M.E., Ivy, J.V., Kalicki, J., Elliott, G., Ries, R.E., Payton, J.E., Westervelt, P., Tomasson, M.H., Watson, M.A., Baty, J., Heath, S., Shannon, W.D., Nagarajan, R., Link, D.C., Walter, M.J., Graubert, T.A., DiPersio, J.F., Wilson, R.K. and Ley, T.J. (2009) 'Recurring mutations found by sequencing an acute myeloid leukemia genome', *N Engl J Med*, 361(11), pp. 1058-66.

Marelli, A.J., Mackie, A.S., Ionescu-Ittu, R., Rahme, E. and Pilote, L. (2007) 'Congenital heart disease in the general population: changing prevalence and age distribution', *Circulation*, 115(2), pp. 163-72.

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*. 437. [Online]. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16056220

Marian, A.J. and Roberts, R. (2001) 'The molecular genetic basis for hypertrophic cardiomyopathy', *J Mol Cell Cardiol*, 33(4), pp. 655-70.

Marino, B. and Digilio, M.C. (2000) 'Congenital heart disease and genetic syndromes: specific correlation between cardiac phenotype and genotype', *Cardiovasc Pathol*, 9(6), pp. 303-15.
Marth, G.T., Yu, F., 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., Altshuler, D., Bustamante, C.D., Clark, A.G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A. and Gibbs, R. (2011) 'The functional spectrum of low-frequency coding variation', *Genome Biol*, 12(9), p. R84.

Mayberry, J.C., Scott, W.A. and Goldberg, S.J. (1990) 'Increased birth prevalence of cardiac defects in Yuma, Arizona', *J Am Coll Cardiol*, 16(7), pp. 1696-700.

McBride, K.L., Riley, M.F., Zender, G.A., Fitzgerald-Butt, S.M., Towbin, J.A., Belmont, J.W. and Cole, S.E. (2008) 'NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling', *Hum Mol Genet*, 17(18), pp. 2886-93.
McDermott, S., O'Neill, A.C., Ridge, C.A. and Dodd, J.D. (2012) 'Investigation of cardiomyopathy using cardiac magnetic resonance imaging part 1: Common phenotypes', *World J Cardiol*, 4(4), pp. 103-11.

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. 2567-74.

McKellar, S.H., Tester, D.J., Yagubyan, M., Majumdar, R., Ackerman, M.J. and Sundt, T.M., 3rd (2007) 'Novel NOTCH1 mutations in patients with bicuspid aortic valve disease and thoracic aortic aneurysms', *J Thorac Cardiovasc Surg*, 134(2), pp. 290-6.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res*, 20(9), pp. 1297-303.

Mi, Y.P., Chau, A.K., Chiu, C.S., Yung, T.C., Lun, K.S. and Cheung, Y.F. (2005) 'Evolution of the management approach for pulmonary atresia with intact ventricular septum', *Heart*, 91(5), pp. 657-63.

Miller, A., Siffel, C., Lu, C., Riehle-Colarusso, T., Frias, J.L. and Correa, A. (2010) 'Long-term survival of infants with atrioventricular septal defects', *J Pediatr*, 156(6), pp. 994-1000.
Miller, M.K., Bang, M.L., Witt, C.C., Labeit, D., Trombitas, C., Watanabe, K., Granzier, H., McElhinny, A.S., Gregorio, C.C. and Labeit, S. (2003) 'The muscle ankyrin repeat proteins: CARP, ankrd2/Arpp and DARP as a family of titin filament-based stress response molecules', *J Mol Biol*, 333(5), pp. 951-64.

Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C. and Devine, S.E. (2011) 'Natural genetic variation caused by small insertions and deletions in the human genome', *Genome Res*, 21(6), pp. 830-9.

Minich, L.L., Atz, A.M., Colan, S.D., Sleeper, L.A., Mital, S., Jaggers, J., Margossian, R., Prakash, A., Li, J.S., Cohen, M.S., Lacro, R.V., Klein, G.L. and Hawkins, J.A. (2010) 'Partial and transitional atrioventricular septal defect outcomes', *Ann Thorac Surg*, 89(2), pp. 530-6.
Minoche, A.E., Dohm, J.C. and Himmelbauer, H. (2011) 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems', *Genome Biol*, 12(11), p. R112.

Miyamoto, Y., Akita, H., Shiga, N., Takai, E., Iwai, C., Mizutani, K., Kawai, H., Takarada, A. and Yokoyama, M. (2001) 'Frequency and clinical characteristics of dilated cardiomyopathy caused by desmin gene mutation in a Japanese population', *Eur Heart J*, 22(24), pp. 2284-9.

Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V. and Cuppen, E. (2010) 'Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries', *Nucleic Acids Res*, 38(10), p. e116.

Morimoto, S., Lu, Q.W., Harada, K., Takahashi-Yanaga, F., Minakami, R., Ohta, M., Sasaguri, T. and Ohtsuki, I. (2002) 'Ca(2+)-desensitizing effect of a deletion mutation Delta K210 in cardiac troponin T that causes familial dilated cardiomyopathy', *Proc Natl Acad Sci U S A*, 99(2), pp. 913-8.

Moulik, M., Vatta, M., Witt, S.H., Arola, A.M., Murphy, R.T., McKenna, W.J., Boriek, A.M., Oka, K., Labeit, S., Bowles, N.E., Arimura, T., Kimura, A. and Towbin, J.A. (2009) 'ANKRD1, the gene encoding cardiac ankyrin repeat protein, is a novel dilated cardiomyopathy gene', *J Am Coll Cardiol*, 54(4), pp. 325-33.

Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., Fennell, T., Banks, E., Ambrogio, L., Cibulskis, K., Kernytsky, A., Gonzalez, E., Rudzicz, N., Engert, J.C., DePristo, M.A., Daly, M.J., Cohen, J.C., Hobbs, H.H., Altshuler, D., Schonfeld, G., Gabriel, S.B., Yue, P. and Kathiresan, S. (2010) 'Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia', *N Engl J Med*, 363(23), pp. 2220-7.
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. 387-9.

Newbury-Ecob, R.A., Leanage, R., Raeburn, J.A. and Young, I.D. (1996) 'Holt-Oram syndrome: a clinical genetic study', *J Med Genet*, 33(4), pp. 300-7.

Ng, C.M., Cheng, A., Myers, L.A., Martinez-Murillo, F., Jie, C., Bedja, D., Gabrielson, K.L., Hausladen, J.M., Mecham, R.P., Judge, D.P. and Dietz, H.C. (2004) 'TGF-beta-dependent pathogenesis of mitral valve prolapse in a mouse model of Marfan syndrome', *J Clin Invest*, 114(11), pp. 1586-92.

Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L. and Venter, J.C. (2008) 'Genetic variation in an individual human exome', *PLoS Genet*, 4(8), p. e1000160.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., Lee, C., Turner, E.H., Smith, J.D., Rieder, M.J., Yoshiura, K., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D.A., Bamshad, M.J. and Shendure, J. (2010a) 'Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome', *Nat Genet*, 42(9), pp. 790-3.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J. and Bamshad, M.J. (2010b) 'Exome sequencing identifies the cause of a mendelian disorder', *Nat Genet*, 42(1), pp. 30-5.

Ng, S.B., Nickerson, D.A., Bamshad, M.J. and Shendure, J. (2010c) 'Massively parallel sequencing and rare disease', *Hum Mol Genet*, 19(R2), pp. R119-24.

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.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. and Barron, A.E. (2011) 'Landscape of next-generation sequencing technologies', *Anal Chem*, 83(12), pp. 4327-41.
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.

Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) 'SSAHA: a fast search method for large DNA databases', *Genome Res*, 11(10), pp. 1725-9.

Norton, N., Li, D., Rieder, M.J., Siegfried, J.D., Rampersaud, E., Zuchner, S., Mangos, S., Gonzalez-Quintana, J., Wang, L., McGee, S., Reiser, J., Martin, E., Nickerson, D.A. and Hershberger, R.E. (2011) 'Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy', *Am J Hum Genet*, 88(3), pp. 273-82.

Nozu, K., Iijima, K., Nozu, Y., Ikegami, E., Imai, T., Fu, X.J., Kaito, H., Nakanishi, K., Yoshikawa, N. and Matsuo, M. (2009) 'A deep intronic mutation in the SLC12A3 gene leads to Gitelman syndrome', *Pediatr Res*, 66(5), pp. 590-3.

Oechslin, E.N., Attenhofer Jost, C.H., Rojas, J.R., Kaufmann, P.A. and Jenni, R. (2000) 'Long-term follow-up of 34 adults with isolated left ventricular noncompaction: a distinct cardiomyopathy with poor prognosis', *J Am Coll Cardiol*, 36(2), pp. 493-500.

Oji, V. and Traupe, H. (2006) 'Ichthyoses: differential diagnosis and molecular genetics', *Eur J Dermatol*, 16(4), pp. 349-59.

Omeri, M.A., Bishop, M., Oakley, C., Bentall, H.H. and Cleland, W.P. (1965) 'The Mitral Valve in Endocardial Cushion Defects', *Br Heart J*, 27, pp. 161-76.

Pagani, F. and Baralle, F.E. (2004) 'Genomic variants in exons and introns: identifying the splicing spoilers', *Nat Rev Genet*, 5(5), pp. 389-96.

Pallavicini, A., Kojic, S., Bean, C., Vainzof, M., Salamon, M., Ievolella, C., Bortoletto, G., Pacchioni, B., Zatz, M., Lanfranchi, G., Faulkner, G. and Valle, G. (2001) 'Characterization of human skeletal muscle Ankrd2', *Biochem Biophys Res Commun*, 285(2), pp. 378-86.
Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011) 'Sequencing technologies and genome sequencing', *J Appl Genet*, 52(4), pp. 413-35.

Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M. and McCombie, W.R. (2011) 'A comparative analysis of exome capture', *Genome Biol*, 12(9), p. R97.
Parmacek, M.S. and Solaro, R.J. (2004) 'Biology of the troponin complex in cardiac myocytes', *Prog Cardiovasc Dis*, 47(3), pp. 159-76.

Parvari, R. and Levitas, A. (2012) 'The Mutations Associated with Dilated Cardiomyopathy', *Biochem Res Int*, 2012, p. 639250.

Parvez, B. and Darbar, D. (2011) 'The "missing" link in atrial fibrillation heritability', *J Electrocardiol*, 44(6), pp. 641-4.

Patel, R.K. and Jain, M. (2012) 'NGS QC Toolkit: a toolkit for quality control of next generation sequencing data', *PLoS One*, 7(2), p. e30619.

Pattnaik, S., Vaidyanathan, S., Pooja, D.G., Deepak, S. and Panda, B. (2012) 'Customisation of the exome data analysis pipeline using a combinatorial approach', *PLoS One*, 7(1), p. e30080.
Paynter, N.P., Chasman, D.I., Pare, G., Buring, J.E., Cook, N.R., Miletich, J.P. and Ridker, P.M. (2010) 'Association between a literature-based genetic risk score and cardiovascular events in women', *JAMA*, 303(7), pp. 631-7.

Pfeffer, G., Elliott, H.R., Griffin, H., Barresi, R., Miller, J., Marsh, J., Evila, A., Vihola, A., Hackman, P., Straub, V., Dick, D.J., Horvath, R., Santibanez-Koref, M., Udd, B. and Chinnery, P.F. (2012) 'Titin mutation segregates with hereditary myopathy with early respiratory failure', *Brain*, 135(Pt 6), pp. 1695-713.

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.

Pyle, A., Griffin, H., Yu-Wai-Man, P., Duff, J., Eglon, G., Pickering-Brown, S., Santibanez-Korev, M., Horvath, R. and Chinnery, P.F. (2012) 'Prominent Sensorimotor Neuropathy Due to SACS Mutations Revealed by Whole-Exome SequencingSensorimotor Neuropathy Due to SACS Mutations', *Arch Neurol*, pp. 1-4.

Qi, J., Zhao, F., Buboltz, A. and Schuster, S.C. (2010) 'inGAP: an integrated next-generation genome analysis pipeline', *Bioinformatics*, 26(1), pp. 127-9.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13, p. 341.
Raju, H., Alberg, C., Sagoo, G.S., Burton, H. and Behr, E.R. (2011) 'Inherited cardiomyopathies', *BMJ*, 343, p. d6966.

Reich, D.E. and Lander, E.S. (2001) 'On the allelic spectrum of human disease', *Trends Genet*, 17(9), pp. 502-10.

Reller, M.D., Strickland, M.J., Riehle-Colarusso, T., Mahle, W.T. and Correa, A. (2008) 'Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005', *J Pediatr*, 153(6), pp. 807-13.

Remenyi, B. and Gentles, T.L. (2012) 'Congenital mitral valve lesions : Correlation between morphology and imaging', *Ann Pediatr Cardiol*, 5(1), pp. 3-12.
Richards, A.A. and Garg, V. (2010) 'Genetics of congenital heart disease', *Curr Cardiol Rev*, 6(2), pp. 91-7.

Rio Frio, T., McGee, T.L., Wade, N.M., Iseli, C., Beckmann, J.S., Berson, E.L. and Rivolta, C. (2009) 'A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance', *Hum Mutat*, 30(9), pp. 1340-7.

Robinson, S.W., Morris, C.D., Goldmuntz, E., Reller, M.D., Jones, M.A., Steiner, R.D. and Maslen, C.L. (2003) 'Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects', *Am J Hum Genet*, 72(4), pp. 1047-52.

Rosamond, W., Flegal, K., Furie, K., Go, A., Greenlund, K., Haase, N., Hailpern, S.M., Ho, M., Howard, V., Kissela, B., Kittner, S., Lloyd-Jones, D., McDermott, M., Meigs, J., Moy, C., Nichol, G., O'Donnell, C., Roger, V., Sorlie, P., Steinberger, J., Thom, T., Wilson, M. and Hong, Y. (2008) 'Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee', *Circulation*, 117(4), pp. e25-146.

Rottbauer, W., Wessels, G., Dahme, T., Just, S., Trano, N., Hassel, D., Burns, C.G., Katus, H.A. and Fishman, M.C. (2006) 'Cardiac myosin light chain-2: a novel essential component of thick-myofilament assembly and contractility of the heart', *Circ Res*, 99(3), pp. 323-31.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D.N., Renwick, A., Seal, S., Ramsay, E., Duarte Sdel, V., Rivas, M.A., Warren-Perry, M., Zachariou, A., Campion-Flora, A., Hanks, S., Murray, A., Pour, N.A., Douglas, J., Gregory, L., Rimmer, A., Walker, N.M., Yang, T.P., Adlard, J.W., Barwell, J., Berg, J., Brady, A.F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Eccles, D., Evans, D.G., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P.J., Paterson, J., Porteous, M., Rogers, M.T., Shanley, S., Walker, L., Gore, M., Houlston, R., Brown, M.A., Caufield, M.J., Deloukas, P., McCarthy, M.I., Todd, J.A., Breast, Ovarian Cancer Susceptibility, C., Wellcome Trust Case Control, C., Turnbull, C., Reis-Filho, J.S., Ashworth, A., Antoniou, A.C., Lord, C.J., Donnelly, P. and Rahman, N. (2013) 'Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer', *Nature*, 493(7432), pp. 406-10.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proc Natl Acad Sci U S A*, 74(12), pp. 5463-7.

Scheper, G.C., van der Knaap, M.S. and Proud, C.G. (2007) 'Translation matters: protein synthesis defects in inherited disease', *Nat Rev Genet*, 8(9), pp. 711-23.

Schonberger, J. and Seidman, C.E. (2001) 'Many roads lead to a broken heart: the genetics of dilated cardiomyopathy', *Am J Hum Genet*, 69(2), pp. 249-60.

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.

Schwarz, J.M., Rodelsperger, C., Schuelke, M. and Seelow, D. (2010) 'MutationTaster evaluates disease-causing potential of sequence alterations', *Nat Methods*, 7(8), pp. 575-6.
Sheffield, V.C., Pierpont, M.E., Nishimura, D., Beck, J.S., Burns, T.L., Berg, M.A., Stone, E.M., Patil, S.R. and Lauer, R.M. (1997) 'Identification of a complex congenital heart defect susceptibility locus by using DNA pooling and shared segment analysis', *Hum Mol Genet*, 6(1), pp. 117-21.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A. and Yu, F. (2010) 'A SNP discovery method to assess variant allele probability from next-generation resequencing data', *Genome Res*, 20(2), pp. 273-80.
Shimada, E., Kinoshita, M. and Murata, K. (2009) 'Expression of cardiac myosin light chain 2 during embryonic heart development in medaka fish, Oryzias latipes, and phylogenetic relationship with other myosin light chains', *Dev Growth Differ*, 51(1), pp. 1-16.
Silversides, C.K., Kiess, M., Beauchesne, L., Bradley, T., Connelly, M., Niwa, K., Mulder, B., Webb, G., Colman, J. and Therrien, J. (2010) 'Canadian Cardiovascular Society 2009 Consensus Conference on the management of adults with congenital heart disease: outflow tract obstruction, coarctation of the aorta, tetralogy of Fallot, Ebstein anomaly and Marfan's syndrome', *Can J Cardiol*, 26(3), pp. e80-97.

Smith, E.N., Chen, W., Kahonen, M., Kettunen, J., Lehtimaki, T., Peltonen, L., Raitakari, O.T., Salem, R.M., Schork, N.J., Shaw, M., Srinivasan, S.R., Topol, E.J., Viikari, J.S., Berenson, G.S. and Murray, S.S. (2010) 'Longitudinal genome-wide association of cardiovascular disease risk factors in the bogalusa heart study', *PLoS Genet*, 6(9).

Smith, K.R., Bromhead, C.J., Hildebrand, M.S., Shearer, A.E., Lockhart, P.J., Najmabadi, H., Leventer, R.J., McGillivray, G., Amor, D.J., Smith, R.J. and Bahlo, M. (2011) 'Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes', *Genome Biol*, 12(9), p. R85.

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., J, O.S., 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. (2012) 'Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease', *Am J Hum Genet*, 91(3), pp. 489-501.

Starfield, M., Hennies, H.C., Jung, M., Jenkins, T., Wienker, T., Hull, P., Spurdle, A., Kuster, W., Ramsay, M. and Reis, A. (1997) 'Localization of the gene causing keratolytic winter erythema to chromosome 8p22-p23, and evidence for a founder effect in South African Afrikaans-speakers', *Am J Hum Genet*, 61(2), pp. 370-8.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S. and Cooper, D.N. (2009) 'The Human Gene Mutation Database: 2008 update', *Genome Med*, 1(1), p. 13.

Stitziel, N.O., Kiezun, A. and Sunyaev, S. (2011) 'Computational and statistical approaches to analyzing variants identified by exome sequencing', *Genome Biol*, 12(9), p. 227.

Sulonen, A.M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A. and Saarela, J. (2011) 'Comparison of solution-based exome capture methods for next generation sequencing', *Genome Biol*, 12(9), p. R94.

Summerell, J., Persuad, V., Miller, C. and Talerman, A. (1968) 'Congenital mitral atresia', *Br Heart J*, 30(2), pp. 249-54.

Syrris, P., Ward, D., Asimaki, A., Evans, A., Sen-Chowdhry, S., Hughes, S.E. and McKenna, W.J. (2007) 'Desmoglein-2 mutations in arrhythmogenic right ventricular cardiomyopathy: a genotype-phenotype characterization of familial disease', *Eur Heart J*, 28(5), pp. 581-8.

Syrris, P., Ward, D., Evans, A., Asimaki, A., Gandjbakhch, E., Sen-Chowdhry, S. and McKenna, W.J. (2006) 'Arrhythmogenic right ventricular dysplasia/cardiomyopathy associated with mutations in the desmosomal gene desmocollin-2', *Am J Hum Genet*, 79(5), pp. 978-84.
Tavtigian, S.V., Deffenbaugh, A.M., Yin, L., Judkins, T., Scholl, T., Samollow, P.B., de Silva, D., Zharkikh, A. and Thomas, A. (2006) 'Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral', *J Med Genet*, 43(4), pp. 295-305.

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

Tesson, F., Sylvius, N., Pilotto, A., Dubosq-Bidot, L., Peuchmaurd, M., Bouchier, C., Benaiche, A., Mangin, L., Charron, P., Gavazzi, A., Tavazzi, L., Arbustini, E. and Komajda, M. (2000) 'Epidemiology of desmin and cardiac actin gene mutations in a european population of dilated cardiomyopathy', *Eur Heart J*, 21(22), pp. 1872-6.

Timson, D.J. (2003) 'Fine tuning the myosin motor: the role of the essential light chain in striated muscle myosin', *Biochimie*, 85(7), pp. 639-45.

Towbin, J.A., Bowles, K.R. and Bowles, N.E. (1999) 'Etiologies of cardiomyopathy and heart failure', *Nat Med*, 5(3), pp. 266-7.

Towbin, J.A., Lowe, A.M., Colan, S.D., Sleeper, L.A., Orav, E.J., Clunie, S., Messere, J., Cox, G.F., Lurie, P.R., Hsu, D., Canter, C., Wilkinson, J.D. and Lipshultz, S.E. (2006) 'Incidence, causes, and outcomes of dilated cardiomyopathy in children', *JAMA*, 296(15), pp. 1867-76.

Trivedi, B., Smith, P.B., Barker, P.C., Jaggers, J., Lodge, A.J. and Kanter, R.J. (2011) 'Arrhythmias in patients with hypoplastic left heart syndrome', *Am Heart J*, 161(1), pp. 138-44.

Trusler, G.A., Yamamoto, N., Williams, W.G., Izukawa, T., Rowe, R.D. and Mustard, W.T. (1976) 'Surgical treatment of pulmonary atresia with intact ventricular septum', *Br Heart J*, 38(9), pp. 957-60.

Tsubata, S., Bowles, K.R., Vatta, M., Zintz, C., Titus, J., Muhonen, L., Bowles, N.E. and Towbin, J.A. (2000) 'Mutations in the human delta-sarcoglycan gene in familial and sporadic dilated cardiomyopathy', *J Clin Invest*, 106(5), pp. 655-62.

Vallania, F.L., Druley, T.E., Ramos, E., Wang, J., Borecki, I., Province, M. and Mitra, R.D. (2010) 'High-throughput discovery of rare insertions and deletions in large cohorts', *Genome Res*, 20(12), pp. 1711-8.

Van der Hauwaert, L.G. and Michaelsson, M. (1971) 'Isolated right ventricular hypoplasia', *Circulation*, 44(3), pp. 466-74.

van der Zwaag, P.A., van Rijsingen, I.A., Asimaki, A., Jongbloed, J.D., van Veldhuisen, D.J., Wiesfeld, A.C., Cox, M.G., van Lochem, L.T., de Boer, R.A., Hofstra, R.M., Christiaans, I., van Spaendonck-Zwarts, K.Y., Lekanne Dit Deprez, R.H., Judge, D.P., Calkins, H., Suurmeijer, A.J., Hauer, R.N., Saffitz, J.E., Wilde, A.A., van den Berg, M.P. and van Tintelen, J.P. (2012) 'Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: evidence supporting the concept of arrhythmogenic cardiomyopathy', *Eur J Heart Fail*.

Villard, E., Duboscq-Bidot, L., Charron, P., Benaiche, A., Conraads, V., Sylvius, N. and Komajda, M. (2005) 'Mutation screening in dilated cardiomyopathy: prominent role of the beta myosin heavy chain gene', *Eur Heart J*, 26(8), pp. 794-803.

Vissers, L.E., 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., Hoischen, A., de Vries, B.B., Brunner, H.G. and Veltman, J.A. (2010) 'A de novo paradigm for mental retardation', *Nat Genet*, 42(12), pp. 1109-12.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X. and Yang, H. (2008) 'The diploid genome sequence of an Asian individual', *Nature*, 456(7218), pp. 60-5.

Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., Jin, X., Jiang, H., Zhang, P., Shen, L., Guo, J.F., Li, N., Li, Y.R., Lei, L.F., Zhou, J., Du, J., Zhou, Y.F., Pan, Q., Wang, J., Li, R.Q. and Tang, B.S. (2010a) 'TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing', *Brain*, 133(Pt 12), pp. 3510-8.

Wang, K., Li, M. and Hakonarson, H. (2010b) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Res*, 38(16), p. e164.
Wang, W., Wei, Z., Lam, T.W. and Wang, J. (2011) 'Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions', *Sci Rep*, 1, p. 55.

Wang, Y., Guo, L., Cai, S.P., Dai, M., Yang, Q., Yu, W., Yan, N., Zhou, X., Fu, J., Guo, X., Han, P., Wang, J. and Liu, X. (2012) 'Exome Sequencing Identifies Compound Heterozygous Mutations in CYP4V2 in a Pedigree with Retinitis Pigmentosa', *PLoS One*, 7(5), p. e33673.
Ware, S.M. and Jefferies, J.L. (2012) 'New Genetic Insights into Congenital Heart Disease', *J Clin Exp Cardiolog*, S8.

Warnes, C.A., Liberthson, R., Danielson, G.K., Dore, A., Harris, L., Hoffman, J.I., Somerville, J., Williams, R.G. and Webb, G.D. (2001) 'Task force 1: the changing profile of congenital heart disease in adult life', *J Am Coll Cardiol*, 37(5), pp. 1170-5.

Way, R.C. (1967) 'Cardiovascular defects and the rubella syndrome', *Can Med Assoc J*, 97(22), pp. 1329-34.

Weary, P.E., Hsu, Y.T., Richardson, D.R., Caravati, C.M. and Wood, B.T. (1969) 'Hereditary sclerosing poikiloderma. Report of two families with an unusual and distinctive genodermatosis', *Arch Dermatol*, 100(4), pp. 413-22.

Weber, J.L., David, D., Heil, J., Fan, Y., Zhao, C. and Marth, G. (2002) 'Human diallelic insertion/deletion polymorphisms', *Am J Hum Genet*, 71(4), pp. 854-62.

Wei, X., Ju, X., Yi, X., Zhu, Q., Qu, N., Liu, T., Chen, Y., Jiang, H., Yang, G., Zhen, R., Lan, Z., Qi, M., Wang, J., Yang, Y., Chu, Y., Li, X., Guang, Y. and Huang, J. (2011) 'Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing', *PLoS One*, 6(12), p. e29500.

Weterman, M.A., Sorrentino, V., Kasher, P.R., Jakobs, M.E., van Engelen, B.G., Fluiter, K., de Wissel, M.B., Sizarov, A., Nurnberg, G., Nurnberg, P., Zelcer, N., Schelhaas, H.J. and Baas, F. (2012) 'A frameshift mutation in LRSAM1 is responsible for a dominant hereditary polyneuropathy', *Hum Mol Genet*, 21(2), pp. 358-70.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. and Rothberg, J.M. (2008) 'The complete genome of an individual by massively parallel DNA sequencing', *Nature*, 452(7189), pp. 872-6.

Wilson, L., Curtis, A., Korenberg, J.R., Schipper, R.D., Allan, L., Chenevix-Trench, G., Stephenson, A., Goodship, J. and Burn, J. (1993) 'A large, dominant pedigree of atrioventricular septal defect (AVSD): exclusion from the Down syndrome critical region on chromosome 21', *Am J Hum Genet*, 53(6), pp. 1262-8.

Wolf, M. and Basson, C.T. (2010) 'The molecular genetics of congenital heart disease: a review of recent developments', *Curr Opin Cardiol*, 25(3), pp. 192-7.

Wren, C., Richmond, S. and Donaldson, L. (2000) 'Temporal variability in birth prevalence of cardiovascular malformations', *Heart*, 83(4), pp. 414-9.

Wu, M.H., Chen, H.C., Lu, C.W., Wang, J.K., Huang, S.C. and Huang, S.K. (2010) 'Prevalence of congenital heart disease at live birth in Taiwan', *J Pediatr*, 156(5), pp. 782-5.

Wu, T.D. and Nacu, S. (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', *Bioinformatics*, 26(7), pp. 873-81.

Wu, T.D. and Watanabe, C.K. (2005) 'GMAP: a genomic mapping and alignment program for mRNA and EST sequences', *Bioinformatics*, 21(9), pp. 1859-75.

Yoskovitz, G., Peled, Y., Gramlich, M., Lahat, H., Resnik-Wolf, H., Feinberg, M.S., Afek, A., Pras, E., Arad, M., Gerull, B. and Freimark, D. (2012) 'A novel titin mutation in adult-onset familial dilated cardiomyopathy', *Am J Cardiol*, 109(11), pp. 1644-50.

Yu, Z., Schneider, C., Boeglin, W.E., Marnett, L.J. and Brash, A.R. (2003) 'The lipoxygenase gene ALOXE3 implicated in skin differentiation encodes a hydroperoxide isomerase', *Proc Natl Acad Sci U S A*, 100(16), pp. 9162-7.

Zarrouk Mahjoub, S., Mehri, S., Ourda, F., Finsterer, J. and Ben Arab, S. (2012) 'Novel m.15434C>A (p.230L>I) Mitochondrial Cytb Gene Missense Mutation Associated with Dilated Cardiomyopathy', *ISRN Cardiol*, 2012, p. 251723.

Zhang, J., Chiodini, R., Badr, A. and Zhang, G. (2011) 'The impact of next-generation sequencing on genomics', *J Genet Genomics*, 38(3), pp. 95-109.

Zhi, D. and Chen, R. (2012) 'Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing', *PLoS One*, 7(2), p. e31358.

Zhu, Y. and Xiong, M. (2012) 'Family-based association studies for next-generation sequencing', *Am J Hum Genet*, 90(6), pp. 1028-45.

Zia, A. and Moses, A.M. (2011) 'Ranking insertion, deletion and nonsense mutations based on their effect on genetic information', *BMC Bioinformatics*, 12, p. 299.

Zot, A.S. and Potter, J.D. (1987) 'Structural aspects of troponin-tropomyosin regulation of skeletal muscle contraction', *Annu Rev Biophys Biophys Chem*, 16, pp. 535-59.

Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S. (2012) 'The mystery of missing heritability: Genetic interactions create phantom heritability', *Proc Natl Acad Sci U S A*, 109(4), pp. 1193-8.

Zwollo, P., Rao, S., Wallin, J.J., Gackstetter, E.R. and Koshland, M.E. (1998) 'The transcription factor NF-kappaB/p50 interacts with the blk gene during B cell activation', *J Biol Chem*, 273(29), pp. 18647-55.