Dissecting the genetic architecture of cardiac disorders through the use of High Throughput Sequencing.

Author:

Cian Murphy

Supervisors:

Dr. Vincent Plagnol

Dr. Pier Lambiase

UCL Genetics Institute

Department of Genetics, Evolution and Environment

October 9, 2016

I, Cian Murphy confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The overriding goal of this thesis was to further refine our understanding of the genetic architecture of cardiomyopathies, Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) and Hypertrophic Cardiomyopathy (HCM). 407 patients with ARVC and 957 with HCM had 41 cardiomyopathy and other putative candidate genes sequenced. By comparing these cohorts against each other and against ethnicity and phenotype matched controls, insights were gained into the role of different types of genetic variants in these conditions.

This in part involved utilising 4500 Whole Exome Sequences (WES) that are part of the UCL-exomes consortium, an in-house dataset that aggregates a diverse set of studies. High throughput DNA sequencing technologies, WES or Whole Genome Sequencing (WGS) are revolutionizing the diagnosis and novel gene discovery for rare disorders. As the field transitions from the early discovery for Mendelian and near Mendelian diseases to more complex and oligo-genic diseases, there is substantial benefit in being able to combine data across studies, performing the type of meta-analysis for cases and controls that have proven to be so successful for Genome-Wide Association Studies (GWAS). However, WGS and WES are substantially more affected by sequencing errors and technical artefacts than genome-wide genotyping arrays. As a consequence, meta-analysis of sequence based association studies are often dominated by spurious associations, which result in technical limitations. Here, we show that it is possible to take advantage of the type of mixed models developed initially to control for population structure in GWAS studies, and apply these ideas to control for technical artefacts.

In an attempt to ascertain the role of CNVs in HCM, these data were examined for the presence

of rare causative CNVs. 12 CNVs were identified from an initial Read Depth approach. 4 of these were subsequently validated by CoNIFER, a bioinformatics method, and Array Comparative Genomic Hybridisation (aCGH): one large deletion in *MYBPC3*, one large deletion in *PDLIM3*, one duplication of the entire *TNNT2* gene and one large duplication in *LMNA*. These results show that the role of CNVs in HCM is small and highlight the efficiency of this two step-strategy.

Acknowledgements

Firstly, I would like to thank my primary supervisor, Dr. Vincent Plagnol. He has provided support from before day one and throughout on statistics to programming and how to do everything else. Apart from technical knowledge, the most lasting thing I feel I have learned is a never ending scepticism of the data, regardless of the pvalue.

My secondary supervisor, Dr. Pier Lambiase offered excellent help on the clinical interpretation of the data alongside general life advice. Dr. Doug Speed of LDAK fame was an immense help in the latter stages of the PhD. I learned a lot more about statistics than I thought I could because of you. Warren, Chris, Lucy, Kitty, Shush, Elvira, Jon, Gareth, Valentina and Julie were some of the best group members I could ask for. From bug catching to the 3.30 coffee routine, you all made it fun to come into UGI. I'll miss you.

I want to thank Daniel, my housemate and fellow PhD student. It was always useful to be able to compare PhDs and brainstorm over dinner. And a hefty amount of tennis and bomberman helped keep me somewhat sane.

The tail end of my PhD was one of the hardest periods of my life. Writing up made me realise how much more I had to and combined with starting medical school resulted in a pretty serious coffee addiction. Noor, thank you so much for putting up with me. Too much help to list, but your humour, patience and food definitely deserve mentions. Best proof reader ever.

I don't need a PhD in genetics to know that I couldn't have gotten here without my parents. Tirelessly supportive, I love you and thank you for giving me the freedom and encouragement to do what

I wanted. Cillian, thank you for helping me with photoshop; there are somethings that even R cannot do it would seem and you were always willing to help me tweak figures until I realised what I wanted. Oisin, you're one of the smartest people I know and as my older brother you were always good at teaching me. Whether or not I wanted to learn.

It was an honour to get to work somewhere the calibre of UCL. Being surrounded by the best researchers I have ever met is the most stimulating environment I could ask for. Lastly, thank you to the British Heart Foundation for being such a good supporter.

Publications arising from this thesis

1. Panagiotis I Sergouniotis, Christina Chakarova, Cian Murphy, Mirjana Becker, Eva Lenassi, Gavin Arno, Monkol Lek, Daniel G Macarthur, Shomi S Bhattacharya, Anthony T Moore, Graham E Holder, Anthony G Robson, Uwe Wolfrum, Andrew R Webster, and Vincent Plagnol. AJHG The American Journal of Human Genetics Biallelic variants in TTLL5 , encoding a tubulin glutamylase , cause retinal dystrophy. American journal of human genetics, 94(5):760769, 2014.

2. Laurence M. Nunn, Luis R. Lopes, Petros Syrris, Cian Murphy, Vincent Plagnol, Eileen Firman, Chrysoula Dalageorgou, Esther Zorio, Diana Domingo, Victoria Murday, Iain Findlay, Alexis Duncan, Gerry Carr-White, Leema Robert, Tela Bueser, Caroline Langman, Simon P Fynn, Martin Goddard, Anne White, Henning Bundgaard, Laura Ferrero-Miliani, Nigel Wheeldon, Simon K. Suvarna, Aliceson O'Beirne, Martin D. Lowe, William J. McKenna, Perry M. Elliott, and Pier D. Lambiase. Diagnostic yield of molecular autopsy in patients with sudden arrhythmic death syndrome using targeted exome sequencing. Europace, page euv285, 2015. ISSN 1099-5129. doi: 10.1093/europace/euv285. U

3. L.R. Lopes, C. Murphy, P. Syrris, C. Dalageorgou, W.J. McKenna, P.M. Elliott, and V. Plagnol. Use of High-throughput Targeted Exome-Sequencing to screen for Copy Number Variation in Hypertrophic Cardiomyopathy. European Journal of Medical Genetics, pages 16, 2015. ISSN 17697212. doi: 10.1016/j.ejmg.2015.10.001.

Acronyms

MYBPC3 Myosin Binding Protein Cardiac 3. 6, 49, 51

RPGR Retinitis Pigmentosa GTPase Regulator. 66

TTLL5 Tubulin Tyrosine Ligase-Like family member 5. 66

1KG 1000 Genome Project. 73

aCGH Array Comparative Genome Hybridisation. 1, 40, 41, 43, 45, 46, 51–62

ArtQ Artefact. 14, 15

ARVC Arrhythmogenic Right Ventricular Cardiomyopathy. 1–3, 6, 7, 19, 21, 24, 28–31, 34, 35, 37, 40, 41, 66, 105, 109, 110, 115

CA Cochran Armitage. 12

CBS Circular Binary Segmentation. 43, 51–62

CGH Comparative Genome Hybridisation. 43

CNV Copy Number Variant. 1, 3, 19, 40, 42, 45, 49, 63, 77

CR Cryptic Relatedness. 65, 77

CV Corrected Variant. 95

DCM Dilated Cardiomyopathy. 41, 64

ECG Electrocardiogram. 41

EMMA Efficient Mixed Model Association. 17

ExAC Exome Aggregation Consortium. 23

FFPE Formalin Fixed Paraffin Embedded. 36

FPL2 Familial Partial Lipodystrophy 2. 64

GATK Genome Analysis Tool Kit https://www.broadinstitute.org/gatk. 17, 18

GC Genomic Control. 8, 12, 13

GIF Genomic Inflation Factor. 98

GWAS Genome Wide Association Study. 2, 16, 81, 126

HCM Hypertrophic Cardiomyopathy. 1, 6, 7, 19, 21, 24, 28–31, 34, 35, 37, 40–42, 63, 64, 66, 109, 110

HLA Human Leukocyte Antigen. 8

HMM Hidden Markov Model. 45

HTS High Throughput Sequencing. 1, 10, 11, 14, 19, 65, 66, 71, 77, 92

HWE Hardy-Weinberg Equilibrium. 81

IBD Inflammatory Bowel Disease. 99, 100

INDEL Insertions-Deletion. 11, 18, 66, 108, 110

LMM Linear Mixed Model. 15, 16, 66, 78, 79, 82, 112

LOF Loss of Function - frameshift and stop-gain or stop loss. 29, 40, 66, 68

MAF Minor Alelle Frequency. 2, 3, 22, 23, 29, 38, 40, 81, 85, 86, 98, 105

MHC Myosin Heavy Chain. 6

MLPA Multiplex Ligation-Dependent Probe Amplification. 63

PC Principal Component. 10, 65, 67, 73, 77–79, 82, 88, 91, 92, 95, 102, 107, 110

PCA Principal Component Analysis. 9, 10, 43, 65–69, 73, 88, 91, 92, 102, 107, 108, 110–112

PCR Polymerase Chain Reaction. 77

PID Primary ImmunoDeficiency. 89

PS Population Stratification. 8, 10, 15, 65, 73, 77, 82, 85, 99, 117, 121, 126

QTL Quantitative Trait Loci. 82

RD Read Depth. 1, 4, 37, 40, 42, 49, 63, 66, 77, 78, 81, 82, 92–94, 111, 112

REML Restricted Maximum Likelihood. 82

RPKM Reads per kilobase per milllion. 43

RRM Realized Relationship Matrix. 16

SADS Sudden Arrhythmic Death Syndrome. 6, 22, 127

SCD Sudden Cardiac Death. 2, 5–7, 19, 21, 28, 36, 37, 84, 114

SKAT Sequence Kernel Association Test. 27, 29

SKAT-O Sequence Kernel Association Test Optimised. 27

SNP Single Nucleotide Polymorphism. 3, 4, 16, 18, 19, 41, 63, 66, 72, 73, 79, 98, 108, 110, 111, 115, 120

SVD Singular Value Decomposition. 43

TK Technical Kinship. 66, 81

UCL-ex UCL Exome Consortium. 1, 17–21, 36, 37, 66, 67, 69, 71–75, 79, 81, 83, 84, 87, 91–95, 98–100, 103, 104, 106, 107, 109, 111–114, 116, 118, 125

VCE Variance Component Estimation. 82

VQSR Variant Quality Score Recalibration. 18

WES Whole Exome Sequencing. 2, 4, 12, 125, 126

WGS Whole Genome Sequencing. 2, 4

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Thesis Overview

This thesis follows a broad theme; that of using High Throughput Sequencing and novel statistical approaches in order to refine our understanding of three of the most common cardiac phenotypes. The rest of this chapter serves as an introduction.

Chapter 2 discusses the analysis of a targeted sequencing experiment of genes related (or thought to be) to HCM and ARVC in two relatively large cohorts of patients with these conditions. I will show how the architecture of these traits, while broadly consistent with the literature, can also differ from published work.

Following that, Chapter 3 builds on the work in [Lopes et al., 2013b] by examining the role of Copy Number Variants (CNVs) in HCM. This is done through a stepwise approach that uses a combination of a RD based method (ExomeDepth) with a Singular Value Decomposition (CoNIFER) followed by validation with Array Comparative Genome Hybridisation (aCGH). RD refers to the number of DNA fragments, reads, that map to a given region during a High Throughput Sequencing (HTS) run.

An in-house consortium of approximately 4500 human whole exome sequences (UCL Exome Consortium) is used as the dataset for Chapter 4. There is substantial benefit in being able to combine data across

studies, performing the type of meta-analysis for cases and controls that have proven to be so successful for Genome Wide Association Study (GWAS). The issue of technical artifacts and genotyping batches has been discussed extensively in the early years of GWAS, and similar concerns are now relevant to Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES). These data are substantially more affected by sequencing errors and technical artifacts than genome-wide genotyping arrays. As a consequence, meta-analysis of sequence based association studies are often dominated by spurious associations, which may result in false positive signals. These issues are usually dealt with by applying stringent quality control cut-offs, which can lead to false negative results. Here, we show that it is possible to take advantage of the type of mixed models developed initially to control for population structure in GWAS studies, and apply these ideas to control for technical artifacts. I show that substantial reduction in the association statistic inflation can be achieved by applying these novel analytical techniques, both for single variant and gene based tests, while preserving the sensitivity of the test. We focus on several cardio-vascular traits (Arrhythmogenic Right Ventricular Cardiomyopathy and Sudden Cardiac Death) to illustrate the ability of these novel methods to produce more interpretable results.

1.2 Key Definitions

Throughout this thesis, some key concepts are referred to. In some cases, they are expanded on further, but here I provide a concise summary for reference.

- MAF - For a given locus, we define Minor Alelle Frequency (MAF) as (the number of minor alleles in the population) / the total number of alleles in the population.
- Effect size - The magnitude of an effect. Can be calculated by subtracting the mean of group 2 from group 1 and dividing by the pooled standard deviation, where pooled standard deviation is $(SD_1 + SD_2)^{0.5}/2$.
- Population Stratification - Refers to the instance where the population in question is not a homogenous population and is instead subject to structure which may or may not be known.

- Missingness - The proportion of missing data. This may be randomly missing or not.
- Genomic Inflation - The genomic inflation factor λ is the ratio of the median of the empirically observed distribution of the test statistic to the expected median. This quantifies the extent of the bulk inflation and the excess false positive rate.
- Single variant and gene based tests - Single variant tests work well for variants that are common (here defined as those with a MAF of $\geq 1\%$) and/or have a large effect size [Li and Leal, 2008b]. For rare and/or low effect size variants, these tests are underpowered and thus have lead to region based testing that assesses the cumulative effect of multiple rare and common variants.

1.3 Structural Variation in the Genome

SNPs are single base pair changes in a DNA sequence and small indels usually refer to variants no greater than 10-20bp. The majority of known disease causing variants are Single Nucleotide Polymorphisms (SNPs) or small indels, which partly reflects the easier challenge to characterise this class of variants in large cohorts. Copy number variants (CNVs) are genetic variants of larger size, either deletion or duplications. CNVs can range in size from kilobases to megabases and can occur spontaneously or be transmitted stably through generations [Feuk et al., 2006].

2010 saw the publication of a 19000 person 8 disease study that identified 3432 CNVs, highlighting the fact they play an important role in many diseases [Craddock et al., 2010]. Before such large scale CNV studies, these loci may have been indirectly tagged by SNPs. Since then, CNVs have been shown to play a role in other diseases, including Schizophrenia [Rees et al., 2014], Duchenne Muscular Dystrophy [Pagnamenta et al., 2011], α-thalassemia [Grimholt et al., 2014] and even short stature [van Duyvenvoorde et al., 2013] as examples. This includes ARVC, which identified a large segregating 122kb deletion in *PKP2* [Li Mura et al., 2013]. At the larger end of the CNV scale, whole chromosomal duplications can occur, leading to conditions such as Trisomy 21 or Turner Syndrome. Large scale characterisation of CNVs is a technical challenge, and therefore much remains to be understood about their role in disease aetiology.

1.4 Exome Sequencing

It is a well established that the cost of the massively parallel sequencing of DNA has plummeted over the recent years at a rate that outpaced Moore's Law [Moore, 1998]. Despite this progress, it is not yet financially viable for mainstream research to routinely sequence the whole genome, a method known as Whole Genome Sequencing (WGS). Therefore, an economical and practical solution is to concentrate efforts on the 1-2% of the genome that are more easily interpretable [Teer and Mullikin, 2010]. This process, Whole Exome Sequencing (WES), covers the exome which consists of all of the known exons across the genome and spans $\sim 30 \times 10^6$ base pairs [Wang et al., 2013]. The basic methodology consists of randomly fragmenting the sample DNA, enrichment of the target exome, exome hybridisation to an array, amplification and finally sequencing [Ng et al., 2009] (Figure 1.1)

Despite its small size, the exome is thought to contain 85% of the variants that cause Mendelian diseases [Wang et al., 2013]. Mendelian refers to genetic phenomena that display complete penetrance (complete correlation between genotype and phenotype) and are caused by a single gene [Marian, 2012]. WES offers the potential to study SNPs and CNVs. Identification of the latter from short-read sequencing offers somewhat more of a challenge than SNPs however as aligning reads to a region with a repetitive sequence can be technically challenging and prone to errors. Three general methods do, however, exist; those that use split reads e.g. [Karakoc et al., 2012], those that take a paired-end approach e.g. [Zeitouni et al., 2010] and finally those that adopt a RD analysis method e.g. [Krumm et al., 2012; Plagnol et al., 2012].

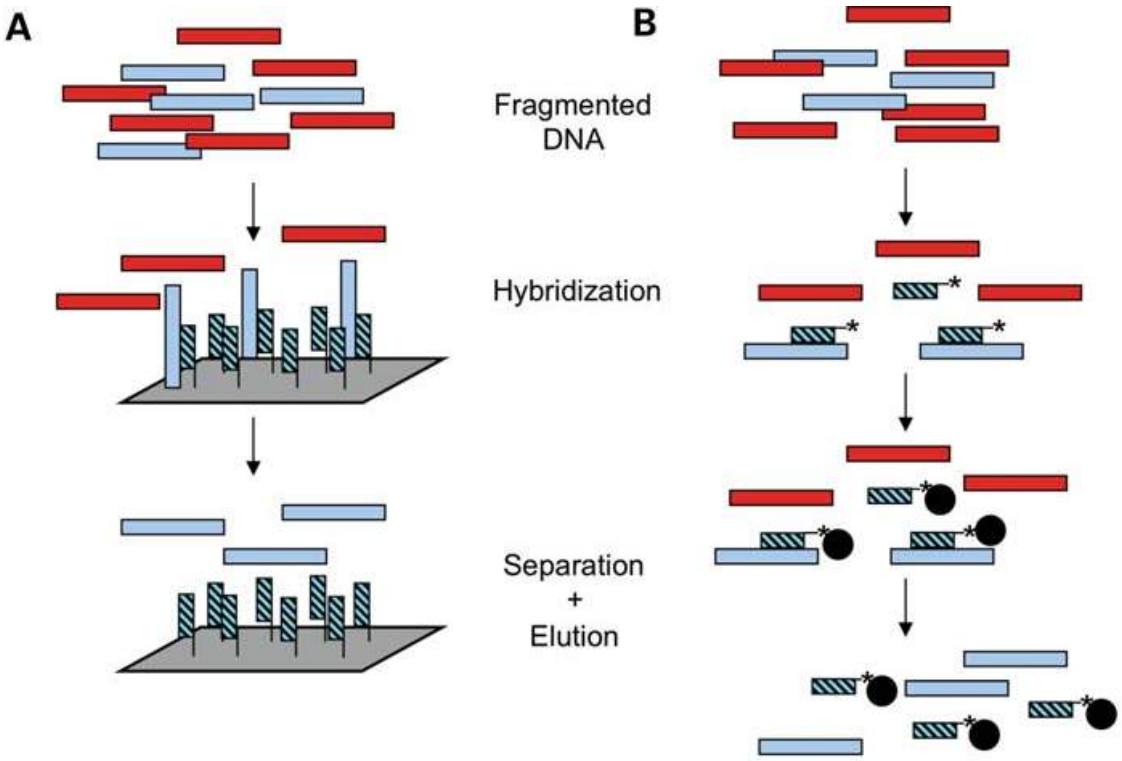


Figure 1.1: Overview of various DNA capture methods. Replicated from [Teer and Mullikin, 2010]. The light blue bar is the target Nucleotide sequence. The red bar represents off-target genomic sequence. (A) An illustration of solid phase hybridisation. Probes (black and light blue) that are complementary to the target sequence are hybridized to a microarray. The fragmented sample DNA is applied and the target sequence binds to the bait probe. The probe is then washed and the fragments are sequenced. (B) Liquid-phase hybridisation. Similar to (A) except the solid substrate (microarray) is replaced with an in-solution reaction that is assisted by biotinylated probes and streptavidin beads.

1.5 Heart Conditions studied in this thesis

1.5.1 Sudden Cardiac Death

Sudden Cardiac Death (SCD) is defined as unexpected natural death that onsets rapidly and has a cardiac origin [Zipes and Wellens, 1998]. Epidemiological studies have shown that the incidence of SCD is $\sim$ 3 to 4 times higher in men than women [Zipes and Wellens, 1998]. While coronary heart disease becomes more frequent with increasing age, SCD in general is a disease of adolescence or early adulthood. Most notably, its effects are exacerbated by physical exercise, leading to a 2.8 fold greater incidence in athletes compared to

non-athletes [Chandra et al., 2013]. SCD is responsible for approximately 500 deaths in England and Wales per annum [Behr et al., 2007]. Clinical screening alone identifies an inherited cardiac condition in 22-53 % of families [Nunn and Lambiase, 2011; Nunn et al., 2015].

In this thesis, SCD refers to the inherited cardiac conditions collectively known as Sudden Arrhythmic Death Syndrome. Sudden Arrhythmic Death Syndrome (SADS) is an umbrella term that describes conditions that fall into two principle categories, structural and electrophysiological. The former consists of Hypertrophic Cardiomyopathy, Arrhythmogenic Right Ventricular Cardiomyopathy and Dilated Cardiomyopathy, the first two of which are examined in detail in this thesis. The latter category includes many conditions, such as Long QT syndrome, Short QT syndrome, Brugada Syndrome, Catecholaminergic polymorphic ventricular tachycardia (CPVT) and Progressive cardiac conduction defect (PCCD) [Millar and Sharma, 2015]. These conditions are all channelopathies in that they interfere with ion transport (and therefore electrical conduction) in the heart.

1.5.2 Hypertrophic Cardiomyopathy

HCM is the most common inherited cardiac disease, with a prevalence of 1/500 in the general population [Efthimiadis et al., 2014]. It is a myocardial form of HCM typified by left ventricular hypertrophy [Ho, 2012] (Figure 1.2). Such hypertrophy, when otherwise unexplained, and greater than 15mm is regarded as the main diagnostic criterion for HCM [Hickey and Rezzadeh, 2013]. Treatment of HCM includes recommendations to reduce the level of physical activity undertaken and may progress to more serious interventions such as β blockers or pacemakers. 50-60% of HCM cases are inherited in an autosomal dominant fashion [Lopes et al., 2013b], caused by mutations in cardiac sarcomeric genes. Z-disc and calcium handling genes are also associated with HCM, but are thought to explain <1% of cases. The sarcomere is the basic unit of muscle that is comprised of myosin thick filaments and actin thin filaments arranged longitudinally [Rahimov and Kunkel, 2013]. The Myosin Heavy Chain (MHC) gene on chromosome 14q1 alone counts for $\sim$ 30 to 50 % of cases, followed by Myosin Binding Protein Cardiac 3 (*MYBPC3*). HCM is characterised by a variable phenotype and incomplete penetrance. As a result of this, family screening of patients with HCM is vital

for effective disease management, while also offering the potential to elucidate the genetic basis.

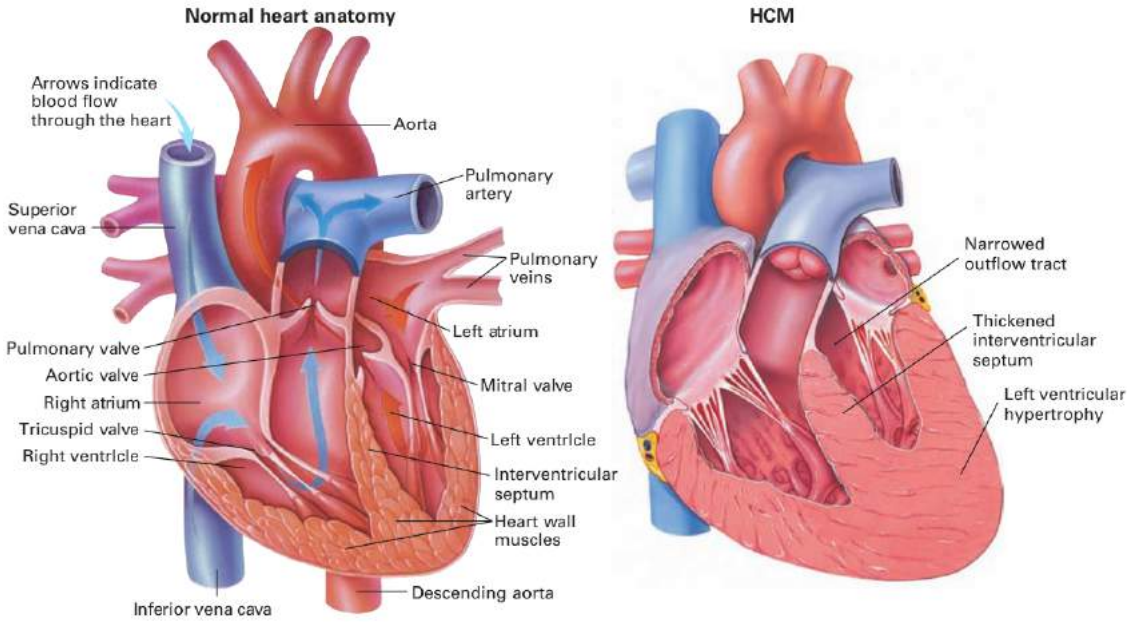


Figure 1.2: Comparison of a normal heart to one with Hypertrophic Cardiomyopathy. Reproduced from [Hickey and Rezzadeh, 2013]

1.5.3 Arrhythmogenic Right Ventricular Cardiomyopathy

ARVC is another inherited cardiomyopathy, primarily affecting the right ventricle [Romero et al., 2013]. It is characterized clinically by fibrofatty replacement, myocardial atrophy, fibrosis, chamber dilation and aneurysm formation [Thiene et al., 1997].ARVC affects men 3 times more than women and has an overall incidence of about 1:5000 [Corrado and Thiene, 2006]. ARVC cases represent approximately 20% of the cases of SCD in the United States [Dalal et al., 2005]. The pathological presentation of ARVC is quite variable, rendering it more difficult to identify its genetic cause than well-defined diseases such as HCM. Nevertheless, some genes have been implicated. The desmosomal gene Desmoplakin (*DSP*) was found to be associated with an autosomal dominant form of ARVC [Rampazzo et al., 2002]. The finding that the genes Junction Plakoglobin (*JUP*) and Plakophilin 2 (*PKP2*) frequently contained mutations in ARVC has suggested that ARVC is a disease of cardiomyocyte junctions [McKoy et al., 2000; Tiso et al., 2001].

1.6 Problems with data interpretation

1.6.1 Population Stratification

Consider this hypothetical situation. One is interested in disease X and knows little about its epidemiology. One therefore decides to collect a cohort of disease samples (cases) from the general population. A prevalence of N% is identified and it is then assumed that this is representative of people as a whole. Furthermore, a particular variant Y (say a SNP) was flagged as associated with the disease. In general, this occurs when it is shown that a variant is significantly over-represented in cases compared to control samples that do not have disease X. There are a number of reasons as to why a variant may indeed have a different frequency between cases and controls. First, Y is a simple false positive and it in actuality has the same frequency in both cases and controls. Secondly, Y is truly disease causing, or is in linkage disequilibrium with the causative allele, and in that case we can mark one more disease off the list of unsolved Mendelian conditions. Finally, Y is neither of the above and is in fact associated with a subpopulation. If X is more common in a particular population, then Y could be associated with their ethnicity rather than with X pathogenesis. These possibilities are summarised in Figure 1.3.

In general, this phenomenon is referred to as Population Stratification (PS). A classic example of a study that failed to implement an adequate control for PS is that in which it was erroneously claimed that there was an association between diabetes and a Human Leukocyte Antigen (HLA) haplotype on a Pima Indian reservation [Knowler et al., 1988]. This association was found because the target population displayed genetic admixture between people of white European and Pima Indian ancestry. PS is thus a source of false positives. When the analysis was restricted to the latter only, the association disappeared [Cardon and Palmer, 2003]. Arguably the easiest solution to PS is to carefully match cases with controls so that their epidemiological background is as similar as possible, except for disease status. With this approach, it can therefore be difficult if not impossible to obtain a sufficiently large and accurate control set. It is not particularly feasible when dealing with rare diseases as the less common the disease of interest is, the larger the required sample size.

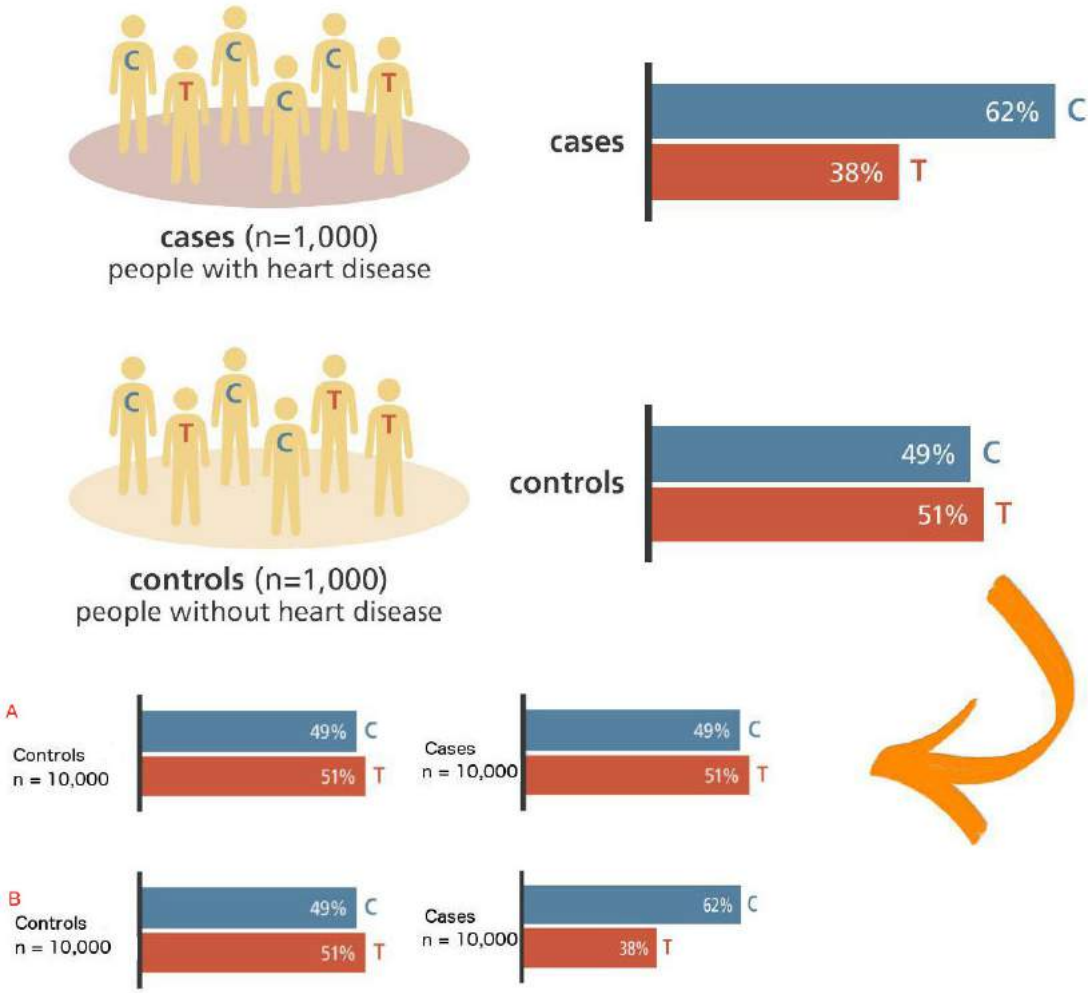


Figure 1.3: The general method of case control studies. In the top panel a SNP frequency is ascertained in 1000 heart disease cases and controls. The arrow indicates two possible explanations for the difference in frequency between cases and controls. A shows how the original finding may be a false positive and the frequency is in fact the same in cases and controls, seen at a larger sample size. B shows cases where the finding is still the same but may be due to it being truly disease associated or caused by factors such as Population Stratification.

An alternative approach is termed Genomic Control (GC). This posits that the χ^2 statistic typically used in case control studies is inflated by some constant factor when there is PS [Devlin and Roeder, 1999; Cardon and Palmer, 2003]. The GC factor is multiplicative and proportional to the level of stratification. It is estimated by examining the unlinked markers on a genome wide level and subsequently used to rescale the χ^2 statistic. GC is popular because it is relatively easy to use but it can be conservative and follows the sometimes unrealistic assumption that all SNPs are affected equally.

Another method, EIGENSTRAT [Price et al., 2006], employs Principal Component Analysis (PCA). PCA was first used in genetics to construct a genetic timeline of how early farming spread across Europe [Menozzi et al., 1978]. PCA calculates the axes that explain the most variation in the data. They are linearly ordered, so the first PC is the axis that explains the most variation. This is a useful technique as it enables visualization of the data in terms of its Principal Components (PCs), also known as Eigenvectors, rather than the traditional X/Y graph approach which is only useful for 2-Dimensional Data. If PS is present in data, the first/top PCs may have axes that have a geographic interpretation [Price et al., 2006]. After PCA, EIGENSTRAT controls for association based on the top few PCs (2-10) before finally computing the ancestry-adjusted association statistics. The top PCs are more likely to reflect large scale differences due to population, rather than causal variation.

There remains a debate as to whether such PCA approaches or model based clustering methods such as STRUCTURE or ADMIXTURE are more useful for association studies [Pritchard et al., 2000; Falush et al., 2003; Patterson et al., 2006; Hoffman, 2013]. Controlling for PS with PCA normally allows you to retain all samples in the study, while STRUCTURE & ADMIXTURE will identify samples that should be removed. STRUCTURE works by using multilocus genotype data to infer population structure in an attempt to probabilistically assign all individuals to one of M (an integer) populations, even where the value of M is unknown. Indeed, Patterson et al. [2006] suggests that a merged system may be used, whereby PCA is used to identify an initial likely value for M before running STRUCTURE. ADMIXTURE is a modification of STRUCTURE; It employs a fast block relaxation scheme using sequential quadratic programming for block updates that translates into a runtime that is nearly equivalent to the faster EIGENSTRAT [Alexander et al., 2009]. Because of this runtime reduction, ADMIXTURE is preferred for studies that have larger sample sizes than that which STRUCTURE could handle.

1.6.2 Other sources of bias

When one considers the potential that massively parallel HTS has to revolutionise population and disease genetics, it should come as no surprise that there exists multiple technologies in this increasingly competitive

market. Two of the most common are those provided by Illumina and Complete Genomics (CG). While the throughput of HTS methods far outpaces that of the traditional Sanger sequencing, their accuracy is less reliable. A study that sequenced the genome of an individual to a coverage of ~76x found that just 88.1% of the ~3.7 million SNPs and Insertions-Deletions (INDELs) were agreed on between Illumina and CG [Lam et al., 2012]. Despite millions of years of evolution, eukaryotes still display a spontaneous mutation rate of $10^{-10} - 10^{-12}$ [Hughes et al., 2005]. So it is not surprising that these technologies are not yet perfect. The confounding that this low concordance could cause is exacerbated by the fact that 1676 genes were found to have platform-specific SNPs. Naively, an argument could be made to remove this problem by simply using one technology for all research. However, their methodologies have some unique advantages. For example, this study found that Illumina reported more errors than CG. Illumina uses a longer read length than that of CG which enables it to sequence regions that CG cannot, such as those that are rich in sequence repeats. This may or may not explain the increased error rate, but it shows that it is beneficial to not discard Illumina nonetheless.

This finding of such discrepancy is far from an isolated incident. The 1000 Genome project (1000G) established to catalogue as much human variation as possible to improve our ability to deduce genotype phenotype correlations [Abecasis et al., 2010]. Quality controls differed between the pilot and intermediate releases and the usage of different technologies led to a false positive rate of 3-17%. This was substantially improved by generating consensus calls from more than one platform, which led to an error rate of 1-4% [Nothnagel et al., 2011]. Therefore, the weight of belief in a candidate variant may be bolstered by it being called by more than one technology. For a lot of researchers however, this is not a practical validation method because of the expense involved. Ultimately, 1000G sequenced 2,504 individuals from 26 different populations [Auton et al., 2015] with this improved methodology.

While technologies that utilise longer read lengths offer a larger, more accurate coverage profile, they do not fully solve these technical biases. The four letters of the DNA alphabet typically pair off in known A-T and G-C couplets. AT bonds consist of two hydrogen bonds while GC pairs use three hydrogen bonds. This fact has a noticeable impact on the performance of PCR and HTS systems, resulting in regions that

contain high numbers of GC pairs (GC rich regions) presenting additional technical difficulties. Illumina has been shown to struggle sequencing GC rich regions, which causes uneven or even a complete lack of coverage [Dohm et al., 2008]. More recently, validation studies have shown that most HTS technologies suffer from some degree of GC bias (Figure 1.4). As this figure shows, when the GC content is close to 50%, then the four technologies examined here perform comparably well. This changes towards either tail (GC rich or GC poor) with the Illumina HiSeq coping significantly better than Life Technologies and even CG.

System updates do not always mean that improvements have been gained in output quality. It has been shown that even more recent versions of the commercially available capture platforms have problems. For example, the WES platforms Agilent (SureSelect v5+UTR), NimbleGen (SeqCap v3+UTR) and Illumina (Nextera Expanded Exome) were compared in a recent study [Chilamakuri et al., 2014; Meienberg et al., 2015]. This showed that Agilent and NimbleGen now perform better than Illumina, despite the latter being the market leader. The latest Agilent platform in particular is the best performer as Nimblegen has a more pronounced GC bias.

Artefactual differences between cases and controls can sometimes exhibit a differential bias that confounds real signal. This was shown in the first phase of the 1958 British birth cohort Diabetes study [Clayton et al., 2005]. As is the norm with genetic first phase studies, the goal was to identify a subset of SNPs from the initial panel that could subsequently be further tested for confirmation on a larger sample. Taqman genotyping was used, which consists of fluorescently ligated PCR primers that target candidate SNPs. The calls for individual genotypes are performed by examining the cluster of fluorescence from cases and controls: in an artefact free world you would expect to see three distinct clusters, a heterozygote cluster flanked by both homozygote pools. However for a range of NS SNPs, the heterozygote clouds for cases and controls were unexpectedly discrete (Figure 1.5). Without correction, this can readily be misconstrued by the clustering algorithm as a false positive. This could be avoided by increasing confidence level required before declaring a call. This is also not ideal as that would both reduce the used variant set and also created 'informative missingness' where missingness is no longer independent of genotype.

To test 2*K contingency tables, such as those seen in genotype studies, the Cochran Armitage (CA)

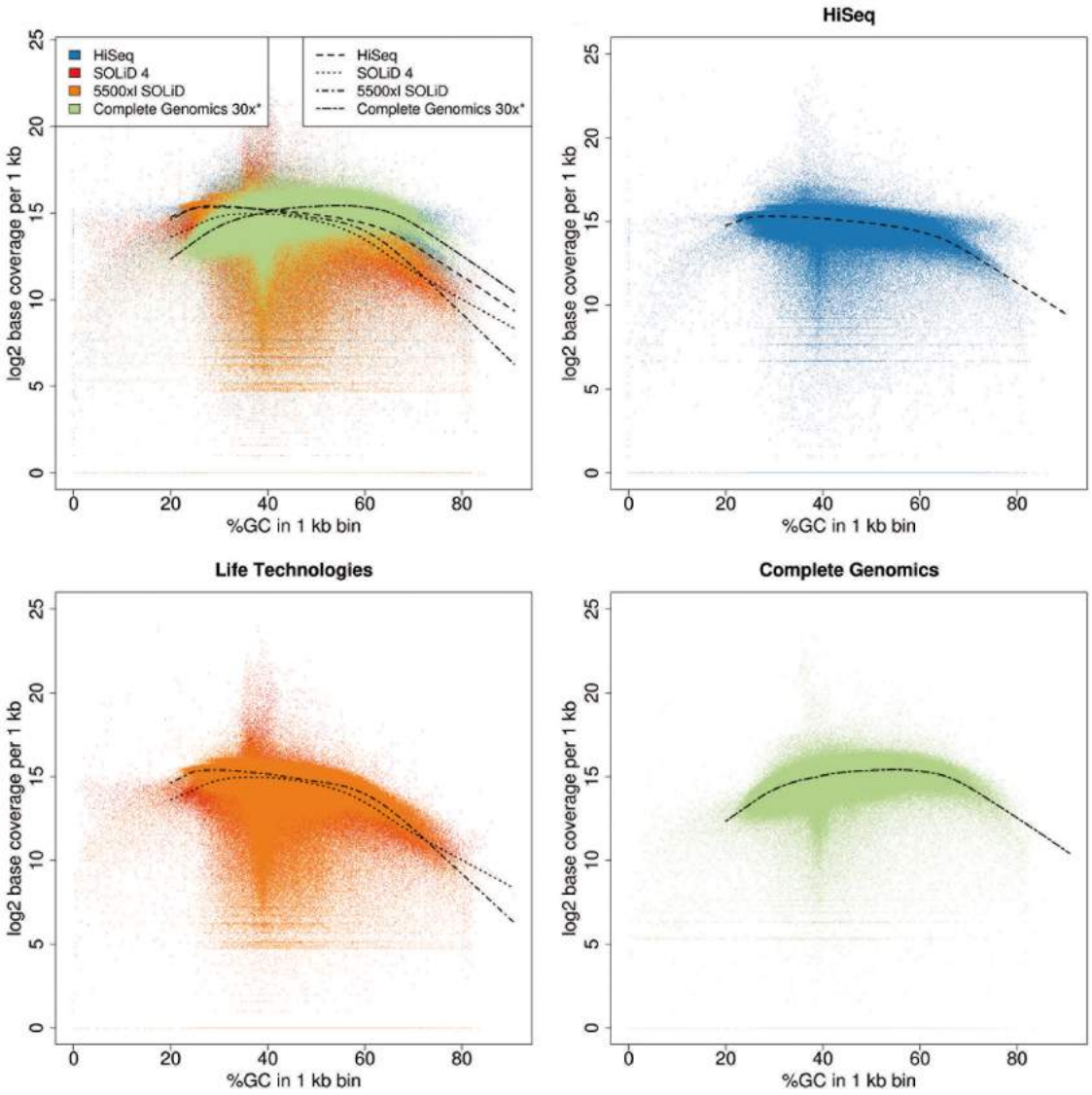


Figure 1.4: GC Bias across four High Throughput Sequencing platforms [Rieber et al., 2013]. Log2 of base coverage in 1 kilobase windows. Top Left panel is a merged picture of the three other panels. A smoothened loess curve was fitted per dataset to show the local coverage.

test is typically used. The null being no association, this will be chi-squared with K-1 degree of freedom. However, Devlin (1999) noticed that substructure in association studies can lead to an overdispersion such that CA is distributed as CA/λ, chi-squared 1df. λ is a constant greater than one that is estimated from a large number of loci throughout the genome. Testing a large number of loci, most of which will be unrelated to the trait of interest, allows one to effectively calculate the background inflation level caused by substructure. This GC is often used to corrected the observed test values by dividing by the estimation of

λ. This method was refined in Clayton (2005) to create a λ that is not constant throughout the genome but depends on regional markers of genotyping accuracy. If no assay based technical bias is present the generalised linear model they implemented reverts λ to just GC.

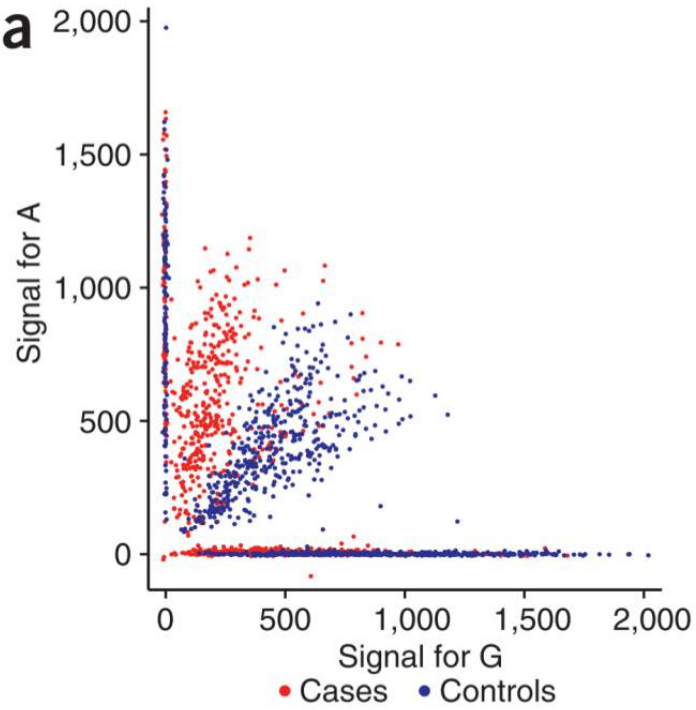


Figure 1.5: Signal intensity plots for the *CD44* SNP rs9666607 from the artefact containing phase 1 diabetes study on the 1958 British birth cohort [Clayton et al., 2005]. The X-axis represents one allele and the Y axis the other. Each dot represents a sample, with those in red cases and blue controls.

The incidence of melanoma, a type of skin cancer, in the Caucasian population has increased by 1.5% annually from 1950 to 2005 [Wang et al., 2009]. The primary cause is solar radiation, with UVA and UVB inducing photoproducts of adjacent pyrimidines which if not adequately remedied can lead to base substitution mutations. The pyrimidines in DNA (Adenine, Thymine and Uracil) would thus be expected to be altered at an increased frequency in melanoma samples. However, in a deep sequencing study of 221 matched melanoma and healthy samples, researchers at the Broad Institute identified a significantly higher rate of purine variants. Figure 1.6 shows that the frequency of this variant substantially increased over time. As this increase was greater than the increase in prevalence, it is suggestive of an altered methodology of

sample preparation, rather than a merely biological cause. Additionally, these were thought to be artefacts as they occurred in a strand specific fashion; The G > T errors were in the first read of the Illumina HiSeq run while the C > A errors were always found in the second read [Costello et al., 2013]. Given that these variants were present in healthy and tumour samples alike and were perfectly correlated with the instrument read order, they were confirmed to be artefacts Artefact (ArtQ).

During preparation for HTS, DNA is randomly fragmented by acoustic and restriction enzyme shearing. The shear force per unit DNA is higher in WES than WGS. This makes it more susceptible to damage, which can manifest as mutations that are erroneously thought to be real signal. This study further identified that some types of DNA storage buffer when exposed to WES methods are responsible for inducing this artefact. To try abrogate this, the best solution would be resequence all samples in ideal buffers and at lower shear forces for WES. However, this method is often impractical and post sequencing corrections are often the only solution.

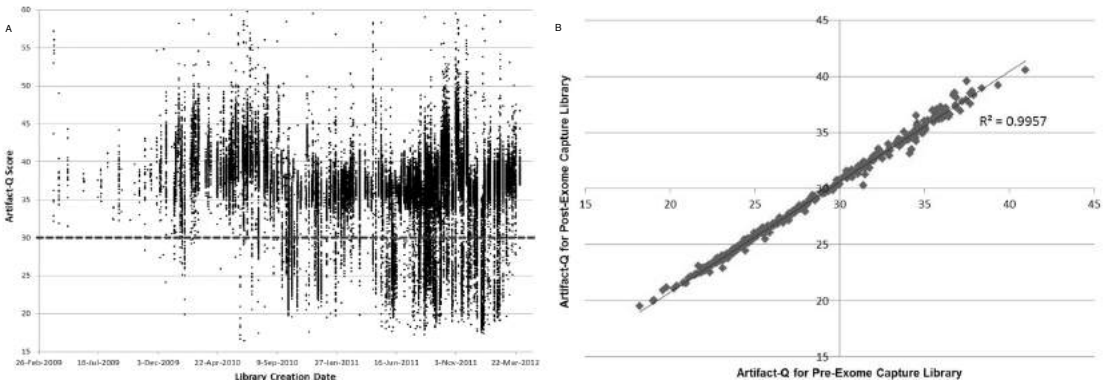


Figure 1.6: A technological ArtQ in a melanoma study. (A) ArtQ prevalence metric by library creation date for the Broad institute's Targeted Capture pipeline. (B) ArtQ for Pre- versus Post-targeted capture. For a set of 370 samples, both the pre- and post-exome enrichment libraries were sequenced. ArtQ was well correlated, indicating that the artifactual base changes had already been introduced before exome capture. Adapted from [Costello et al., 2013].

1.7 Linear Mixed Models

Linear Mixed Models (LMMs) extend the standard Linear Model (Equation 1.1) by adding random effects. They have been used to control for PS alongside methods such as EIGENSTRAT and ADMIXTURE [Zhang

et al., 2010]. LMMs have the form shown in Equation 1.2.

$$Y = Z\alpha + X_j\beta_j + \quad e \quad with \;\; e \sim N(0, I\sigma_e^2) \tag{1.1}$$

$$Y = Z\alpha + X_j\beta_j + \quad g \quad + e \quad with \;\; g \sim N(0, K\sigma_g^2) \tag{1.2}$$

where

- Y is phenotype
- Z is a matrix of covariates
- α is Z's fixed effects
- X_j are the SNPs for SNP j
- β_j is the effect sizes of SNP j
- I is an identity matrix
- e is environmental noise
- and g is a random effect

As with the standard Linear Model, it is necessary to solve Equation 1.2 for each SNP in turn. LMMs can control for multiple types of confounders simultaneously. While this strength is an advantage over these other methods, it has traditionally been such a computationally intensive approach as to be infeasible for GWASs that studied many thousands of markers across thousands of samples. When applied to genetics, LMMs control for confounders by introducing a random effect with correlation structure specified by a "kinship matrix", which measures the genetic similarity between pairs of individuals. This kinship matrix has been estimated with different methods, such as the Realized Relationship Matrix (RRM) [Hayes et al., 2009], an Identity by Descent Approach [de Roos et al., 2009] or by sampling a small set of markers [Lippert et al., 2011]. The last of these has been implemented in the software FaST-LMM.

The LMM log likelihood of Y given X (N x d) which includes the covariates, the SNP and a column of ones as a bias offset can be written as per Equation 1.3. A LMM with a SNP based RRM and without fixed effects is equivalent to a linear regression of the SNPs on the phenotype, with weights integrated over independent Normal distributions having the same variance [Hayes et al., 2009; Lippert et al., 2011]. By replacing K with its spectral decomposition, $K = USU^T$, and by defining δ as σ_g^2/σ_e, one can eventually view this as the linear regression equation (Equation 1.4).

$$likelihood(Y|Data) = Normal(Y|Z\alpha, \sigma_g^2 K + \sigma_e^2 I) \tag{1.3}$$

where

- Normal(Y|,b) denotes a normal distribution with mean a and covariance matrix b

The key to solving Equation 1.2 is determining δ, that is the ratio of the residual variance to the genetic variance. Solving δ naively for each SNP is very computationally intensive, so early implementations such as Efficient Mixed Model Association (EMMA) provided an approximate method which instead solved δ once under the null model, then used this value when testing each SNP. FaST-LMM improves on the algorithm EMMA by reducing the required frequency of Spectral Decompositions from once per SNP to just once [Kang et al., 2008]. It does this with an exact method, by realising that δ can be found rapidly for each SNP after first performing a decomposition of the kinship matrix. FaST-LMM therefore has a runtime and memory footprint that is linear in the number of individuals, making it amenable to data the scale of the UCL-ex consortium.

$$likelihood(Y|Data) = Normal(U^T Y|U^T Z\alpha, \sigma_e^2(\delta S + I)) \tag{1.4}$$

1.8 Bioinformatics - the Genome Analysis Toolkit

1.8.1 Unified Genotyper pipeline with GATK

Raw FASTA files in FASTQ format were aligned to the HG19 reference genome using Novoalign version 2.08.03. Duplicate reads were marked using Picard tools MarkDuplicates.

Until early 2014, and for all the analyses presented in this report, all variants were called using the Unified Genotyper module of the Genome Analysis Tool Kit https://www.broadinstitute.org/gatk (GATK). BAM files were reduced using the GATK ReduceReads module and calling was performed jointly for all samples using GATK version 2.8.1.

1.8.2 Haplotype caller pipeline

Starting in January 2014, calling was performed using the haplotype caller module of GATK, creating gVCF formatted files for each sample. The individual gVCF files were combined into combined gVCF containing 100 samples each. The final variant calling was performed using the GATK "GenotypegVCFs" module jointly for all cases and controls. This process is still being tested. However, preliminary results are very positive. In particular, the computational burden is substantially reduced by the use of this new calling strategy.

1.8.3 VQSR

Variant filtering is central to the methodology presented in this report. The issue of filtering low quality variants has been flagged by all variant calling algorithms, including GATK and `Samtools`. Traditional methods used to flag variants of low quality examined their context, for example, the number of reads covering the region, how many reads cover each allele or the proportion of reads in forward and reverse orientation. Such values were then used to set a threshold and discard variants thus deemed unsatisfactory. These methods are easy to implement but potentially suffer from their crudeness by being too stringent. For UCL-ex data analysis, we followed the best practices as described byGATK to apply the VQSR steps.

Briefly, a set of established summary statistics are computed for at a single variant level. A multi-dimensional mixture model is then fitted to these summary statistics, which allows the computation of a likelihood score for each variant. The further away the summary statistics are from the centers of the Gaussian mixture, the lower the likelihood will be. A training set of established variants is then used, and a likelihood threshold is then set such that a set fraction (typically about 99%) of these established variants passes the threshold. This likelihood threshold is then applied to the dataset as a whole, and variants above that threshold receive a PASS flag. Variants below that threshold are annotated with the "tranche" information that summarizes how far away the summary statistics are from the acceptance threshold. For this report, PASS variants as well as SNPs and INDELs in the top likelihood tranche were included for subsequent analyses.

1.9 Motivation and Aims

The overriding goal of this thesis was to further refine our understanding of the genetic architecture of cardiomyopathies that cause SCD, ARVC and HCM. Several obstacles complicate this aim:

- As discussed already, these conditions display varying levels of penetrance. This makes their analysis more difficult than simple Mendelian conditions.
- The variable phenotypes of these conditions raises the possibility that they are in fact not single conditions and may represent overlapping syndromes. This would further complicate matters as it will weaken any associations found.
- Their relative rarity in the general population means it is not straightforward to establish a cohort with a large number of samples. This limits the possible statistical power.
- To achieve statistical power more samples are needed when studying rare conditions than common ones. This led to the creation of UCL-ex, an in-house collaboration pooling some 4500 whole exomes from cohorts with various rare diseases. This data comes from many different sources with widely disparate results (e.g. in terms of variant call rates and read depth).

Theses obstacles were approached in three distinct methods. Chapter 2 looks at a targeted sequencing approach of these three cohorts. 59 SCD patients are sequenced, using a targeted sequencing approach of known or putative candidate genes, in an attempt to perform a "molecular autopsy" that has an informative diagnostic yield. This is followed by a case control analysis on a targeted sequencing panel of both a HCM and ARVC cohort.

Secondly, Chapter 3 builds on previously published work that examined the role of SNPs in HCM [Lopes et al., 2013b]. It does so by using a three pronged approach to ascertain the role, if any, of CNVs in HCM pathogenesis.

As mentioned already, HTS data can suffer from artefacts derived from many sources. These can be more apparent in a dataset such as UCL-ex where samples from multiple sources are pooled and all have rare diseases. Chapter 4 is devoted to an attempt to create a novel statistical model that adapts classical methods from population genetics to try solve this problem. This is developed and then applied to all the phenotypes in UCL-ex.

Chapter 2

Elucidating the genetic architecture of HCM and ARVC

2.1 Introduction

This chapter examines three cardiomyopathies, Sudden Cardiac Death (SCD), Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) and Hypertrophic Cardiomyopathy (HCM).

59 SCD patients are sequenced, using a targeted sequencing approach of 135 known or putative candidate genes, in an attempt to perform a "molecular autopsy" that has an informative diagnostic yield. Non-synonymous, loss-of-function, and splice-site variants with a minor allele frequency $\leq 0.02\%$ in the NHLBI exome sequencing project and an internal set of control exomes were prioritized for analysis followed by $\leq 0.5\%$ frequency threshold secondary analysis. This initial part was done by others, but I performed the control selection by PCA and the subsequent case control analysis.

This is followed by a case control analysis on a targeted sequencing panel of both a HCM and ARVC cohort. These cohorts were compared against the population controls of UCL-ex to identify novel associations. Additionally, they were compared against other in a bit to identify if this approach is useful in refining our understanding of these somewhat similar conditions. Dr. Pier Lambiase and Dr Petros Syrris

were involved in the sample collection and Dr. Vincent Plagnol performed the sample genome alignment and variant calling. I performed the variant QC, and case control analysis.

2.2 Methods & Results

2.2.1 Molecular Autopsy of a Sudden Arrhythmic Death Syndrome cohort

The cohort analysed here consisted of families referred to specialised cardiovascular centres in seven European centres. Recruited families had a proband who suffered from SADS and was aged between 1-55 with no cause death identified at post-mortem. This study complied with the Declaration of Helsinki and a joint University College London and University College London Hospital Research Ethics committee application. As part of this, the families were offered clinical screening for inherited channelopathies and cardiomyopathies using a standard protocol [Nunn and Lambiase, 2011]. This included an outpatient consultation and resting and exercise electrocardiogram and ajmaline challenge if Brugada syndrome was suspected or was how the proband died or if every other investigation was normal.

90 deceased probands met these initial criteria. 28 were rejected because of DNA quality and/or quantity issues. The next of kin refused consent in 3 additional cases. In total, the DNA from 59 SADS victims (mean age 25, range:1-51) was isolated [Nunn et al., 2016]. The clinical characteristics of these remaining probands are summarised in Figure 2.1. 39/59 patients had structurally normal hearts and 20 had subtle structural abnormalities that were detected post-mortem. Targeted exome sequencing of 135 genes associated with cardiomyopathy or ion channelopathies was performed on the Illumina HiSeq2000 platform (The full list of candidate genes is in Tables 1 to 3 in the Appendix). Variants that were non-synonymous, loss of function or splice site variants with a MAF of $\leq 0.02\%$ in the NHLBI set of 6500 Exomes and the internal control set were prioritised for analysis. Both of these control datasets were filtered to ethnically match the Caucasian cases. The secondary analysis examined variants that had a MAF of $\leq 0.5\%$. Applying this filter yielded 80 candidate coding variants, a mean of 1.36 variants per proband. The variants deemed most likely to be causative based on the gene they are located in and the particular affect that had are listed

in Table 2.1.

Additionally, data from the Exome Aggregation Consortium (ExAC) was used. ExAC is a publically available collection of 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The large size of this dataset meant that it was deemed the most accurate determinant of factors such as MAF. One caveat is the fact that it only became available when this thesis was already at a late stage so it was not possible to retroactively use it in all cases. Despite this, this work gave insight into the clinical utility of a molecular autopsy of sudden cardiac death.

n	59
Mean age (range)	25.3 (1–51) years
Sex	76% male
Circumstances of death	
Daily activities	37%
Sleep/at rest	36%
Exercise	17%
Acoustic stress	2%
Unknown/not recorded	8%
Previous symptoms	20.3%
Syncope	8
Diagnosis of epilepsy	2
Aborted cot death	1
Chest pain	1
Previous 12 lead ECG	13.6%
No specific finding	5
Inferior J point elevation	1
Non-specific intra-ventricular conduction delay	1
Anterior early repolarisation changes not diagnostic of Brugada phenotype	1
Family history of sudden cardiac death	15.3%
Age at death <50 years	5
No age specified	4

Figure 2.1: Summary of proband characteristics in the SADS Molecular Autopsy Study. Reproduced with permission from [Nunn et al., 2016].

AgeAtDeath(yrs)	Sex	CircumstanceDeath	Gene(Disease)	AminoAcidChange	UCLex-MAF(2867ex)	NHLBI(6500ex)	ExAC(60706ex)
4	F	Sleep	*SCN5A*(LQT3/BrS)	R1623Q	0	0	0
6	M	Sleep	*SCN5A*](LQT3/BrS)	V411M	0	0	0
26	F	Phone call	*RyR2*(CPVT)	N1551S	0	0	0.034
18	F	DailyActivities	*TTN*(DCM/HCM)	E23106X	0	0	0.00083
32	M	Sleep	*GJA5*(Familial-AF)	Y197X	0	0	0
39	M	DailyActivities	*MYOT*(LGMD)	Q453X	0	0	0.00165
44	M	DailyActivities	*DSC2*(ARVC)	S868F	0	0.0077	0.0058
23	M	DailyActivities	*CACNA1C*(BrS)	P817S	0.127	0.33	0.0194
1	M	DailyActivities	*LMNA*(DCM)	R644C	0.1385	0.1	0.121
22	F	Sleep	*RANGRF* (BrS)	E61X	0.2646	0.42	0.3947
11	M	Exercise	*CACNA2D1* (BrS)	S709N	0.22	0.37	0.2677
33	M	DailyActivities	*ANK2*(LQT)	E1837K	0.29	0.31	0.267
27	M	DailyActivities	*KCNH2*(LQT)	P347S	0.16	0.0496	0.1293
41	M	DailyActivities	*MYPN*(HCM)	Y20C	0.36	0.092	0.091
28	M	Exercise	*RBM20*(DCM)	E1125K	0.34	0.37	0.37
14	M	Exercise	*DSP*(ARVC)	A2294G	0.12	0.23	0.085
34	M	Sleep	*CACNA1C*(BrS)	G37R	0.3211	0.23	0.74

Table 2.1: Sudden Cardiac Death Molecular Autopsy variants. The first seven are very rare variants (Minor Allele Frequency [MAF] of $\leq .02\%$) in NHLBI and the UCL-exome consortium control set. The last ten are deemed quite rare, with a MAF of $\geq 0.02\%$ & $\leq 0.5\%$

2.2.2 ARVC and HCM case control analysis

As described in Section 1.8.2, the best practises guide from the GATK endorses a joint calling procedure as it has a lower artefact rate than more traditional single sample calling. This method was implemented to more adequately integrate the 407 ARVC and the 955 HCM samples with the 3587 UCL-ex controls. The genes sequenced for this analysis are listed in Table 4 in the Appendix. A case control analysis was then performed, at a single variant and gene level. In total, 9206 variants were tested. The most significantly associated SNPs for ARVC and HCM are listed in Tables 2.2 and 2.3. Given the population prevalence of these conditions (1/500 for HCM and 1/5000 for ARVC) and the varying penetrance of the causative variants, this study was underpowered for rare variants of small effects. For example, if we calculate power for ARVC as it is the rarer condition, and assume:

- A disease prevalence of 1/5000 (0.0002)
- A risk variant population frequency of 0.0002 (A typical value of the most significant variants)
- A heterozygote relative risk of 5
- A homozygote relative risk of 10

rsID	SNP	HUGO	Fisher	OR	ARVC.maf	HCM.maf	Ctrl.maf	nb.cases	nb.ctrls	ESP6500
rs193922674	c.2014C>G	*PKP2*	1.755E-12	6.0732E+01	1.719E-02	5.241E-04	2.882E-04	407	3469	0
NA	c.1957_1963del	*PKP2*	1.17E-08	Inf	9.8280E-03	0	0	407	3567	0
NA	c.1962G>C	*PKP2*	1.18776E-08	Inf	9.8280E-03	0	0	407	3561	0
rs111517471	c.1965C>T	*PKP2*	1.3171E-07	Inf	8.59951E-03	0	0	407	3493	0
rs191564916	c.1003A>G	*DSG2*	3.4655E-07	1.8701E+01	1.22850E-02	5.2410E-04	6.6401E-04	407	3012	0
rs2230234	c.877A>G	*DSG2*	1.2970E-06	1.7921E+00	1.28993E-01	7.6109E-02	7.6317E-02	407	3302	0
rs72648971	c.19389G>A	*TTN*	3.7842E-05	3.8469E+01	7.37101E-03	5.2410E-04	1.9267E-04	407	2595	0
rs72648212	c.53910C>A	*TTN*	7.7010E-05	2.1493E+01	7.37101E-03	5.2410E-04	3.4506E-04	407	2898	0
rs72648909	c.14806T>A	*TTN*	8.2087E-05	1.7216E+01	7.37101E-03	5.2465E-04	4.3078E-04	407	3482	0

Table 2.2: ARVC Single Variant Results for variants with a pvalue of $\leq 1 * 10^{-4}$. rsID is the reference SNP cluster ID. SNP details the position of the tested variant (hg19). Gene is the Ensembl name while HUGO is the HUGO ID. Fisher is the pvalue from Fisher's exact test. OR is the Odds Ratio. ARVC.maf is the variant MAF in ARVC, HCM.maf is its MAF in HCM and Ctrl.maf is the MAF in the controls. nb.cases is the number of ARVC samples called and nb.ctrls is the number of controls called. ESP6500 is the variant MAF in the Exome Sequencing Project.

rsID	SNP	HUGO	Fisher	OR	ARVC.maf	HCM.maf	Ctrl.maf	nb.cases	nb.ctrls	ESP6500
rs375882485	c.1504C>T	*MYBPC3*	1.66196E-12	Inf	0	9.77199E-03	0	921	3220	0
NA	c.2296A>C	*MYBPC3*	2.21412E-09	1.56197E+01	4.88599E-03	1.47569E-02	9.57121E-04	576	2612	0
rs397515916	g.15129A>T	*MYBPC3*	3.79135E-09	2.94978E+01	0	1.01626E-02	3.47826E-04	738	2875	0
rs397516074	c.772G>A	*MYBPC3*	3.05968E-07	Inf	0	5.69476E-03	0	878	3049	0
rs35141404	c.90G>A	*RBM20*	4.39110E-07	1.70479E+00	1.98113E-01	2.73973E-01	1.81210E-01	292	1719	0
NA	c.2374T>C-C	*MYBPC3*	4.78231E-07	Inf	0	5.70776E-03	0	876	2871	0
rs11998271	c.12561C>T	*PLEC*	1.15E-06	2.68031E+00	7.38916E-03	2.58811E-02	9.81308E-03	908	3210	0
rs2340917	c.536T>C	*TMEM43*	1.31346E-06	1.32295E+00	3.18766E-01	3.70074E-01	3.07499E-01	812	3587	0
rs62642469	c.5946G>A	*PLEC*	4.94547E-06	3.21987E+00	2.69542E-03	2.29358E-02	7.23534E-03	654	2626	0
rs72648911	c.15447C>T	*TTN*	8.70525E-06	4.21351E+00	3.68550E-03	1.15425E-02	2.76328E-03	953	3257	0
NA	c.3227A>G	*MYBPC3*	3.58442E-05	Inf	0	4.22195E-03	0	829	2744	0
rs371898076	c.1988G>A	*MYH7*	8.61522E-05	Inf	0	3.14795E-03	0	953	3577	0
NA	c.1063G>A	*MYH7*	8.62942E-05	Inf	0	3.15126E-03	0	952	3572	0
rs35775257	c.9972C>T	*PLEC*	9.24776E-05	3.02707E+00	2.50000E-03	1.44509E-02	4.81965E-03	865	3216	0
rs61233923	c.15252T>C	*TTN*	9.66601E-05	5.30623E+00	1.24069E-03	7.89474E-03	1.49701E-03	950	2672	0

Table 2.3: HCM Single Variant Results for variants with a pvalue of $\leq 1 * 10^{-4}$. rsID is the reference SNP cluster ID. SNP details the position of the tested variant (hg19). Gene is the Ensembl name while HUGO is the HUGO ID. Fisher is the pvalue from Fisher's exact test. OR is the Odds Ratio. ARVC.maf is the variant MAF in ARVC, HCM.maf is its MAF in HCM and Ctrl.maf is the MAF in the controls. nb.cases is the number of HCM samples called and nb.ctrls is the number of controls called. ESP6500 is the variant MAF in the Exome Sequencing Project.

then we would need 1612 cases for 80% power of detecting a real causal variant at an Alpha of 0.1, or 4454 cases at 0.001. This was calculated with the power calculator at http://pngu.mgh.harvard.edu/ purcell/cgi-bin/cc2k.cgi. This low power estimate is in agreement with recent literature highlighting the difficulty of rare variant association studies [Auer et al., 2015].

To increase power, gene-based Fisher, χ^2, Sequence Kernel Association Test (SKAT) and Sequence Kernel Association Test Optimised (SKAT-O) p-values were also calculated (Tables 2.4 and 2.5). The SKAT is a supervised method that performs regressions for each variant within a given region. It differs from burden tests in that it does not upweigh rare variants or assume that pathogenicity increases inversely to variant frequency. The C-Alpha test also allows for varying directions of effects and is essentially a simple version of SKAT where the outcome is binary and no covariates are included. For a dichotomous phenotype, such as case control status, consider the logistic model Equation 2.1.

SKAT has more power than burden tests when variants either have variable effect sizes or effects in different directions, i.e. some SNPs in a gene can be protective and some deleterious [Wu et al., 2011]. However, the inverse is also true; in a scenario where all variants in a set have a unidirectional effect, a burden test will outperform SKAT. SKAT-Optimal (SKAT-O) retains power in either scenario by using an adaptive kernel that follows a multivariate distribution with exchangeable correlation structure [Lee et al., 2012]. For a given set, if the variants effects are uncorrelated, it is effectively a SKAT test, while reducing to a burden test if the effects are unidirectional. Additionally, SKAT can handle singletons by collapsing those with the same directionality of effect into a single value and combining this with the other variants in the region. In general therefore, SKAT-O is more accurate than SKAT as assumptions about variants effects' based on criteria such as predicted function are less important.

$$logitP(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}' X_i + \boldsymbol{\beta}' G_i + \epsilon_i \tag{2.1}$$

where

- y is a binary phenotype vector

- α_0 is an intercept term
- $\boldsymbol{\alpha}$ is the vector of regression coefficients for covariates
- β is a vector of regression coefficients for the p variants in the region.

A lack of statistical power, particularly for ARVC, hindered the ability to identify variants or genes of weak effect here. Additionally, a targeted gene panel of limited size offers little potential to find real novel insight. The findings discussed here do support the literature.

2.2.3 Examining the veracity of candidate gene lists

Table 4 lists the genes with the strongest support for involvement in HCM (16 genes) and ARVC (12 genes). A minority of significantly associated SNPs (here defined as those with a Fisher pvalue of ≤ 0.0001) were seen in candidate genes for HCM (8/53), while 21/28 of the top ARVC SNPs were in candidate genes. HCM Candidate genes *TNNI3,TNNT2,TPM1,MYL2,MYL3,ACTC1,CSRP3,ACTN2,MYH6,TCAP,TNNC1,PLN,MYOZ2,NEXN* did not contain significant SNPs; similarly in the ARVC analysis *CTNNA3,DES,DSC2,DSP,JUP,LMNA,TGFB3,PLN* SNPs failed to reach significance. Loss of function variants were defined as those predicted to be exonic splicing, stopgain/stoploss or frameshift. Tables 2.6 and 2.7 show that *MYBPC3* contains 61% of such variants in HCM while *PKP2* contains 27% of the LOF variants seen in the ARVC samples. Figure 2.2 shows the range of pvalues from the SKAT test for pooled variants and whether or not the fact that the genes are currently candidate genes is a good indicator of the statistical significance that they reach.

As the gene encoding the largest protein in the human body it is unsurprising that variants within Titin have been linked to a number of conditions [Brun et al., 2014; Herman et al., 2012; De Cid et al., 2015]. Because of this clinical importance much work has been done to elucidate its role in these varying pathologies. In terms of SCD, it has been shown that variants in different protein domains are associated with particular types of SCD, as seen in Figure 2.3.

In general, the risk allele for qualitative traits is the minor allele [Park et al., 2011]. One possible definition of effect size is the coefficient (β) for a SNP when it is modeled in a logistic regression against the

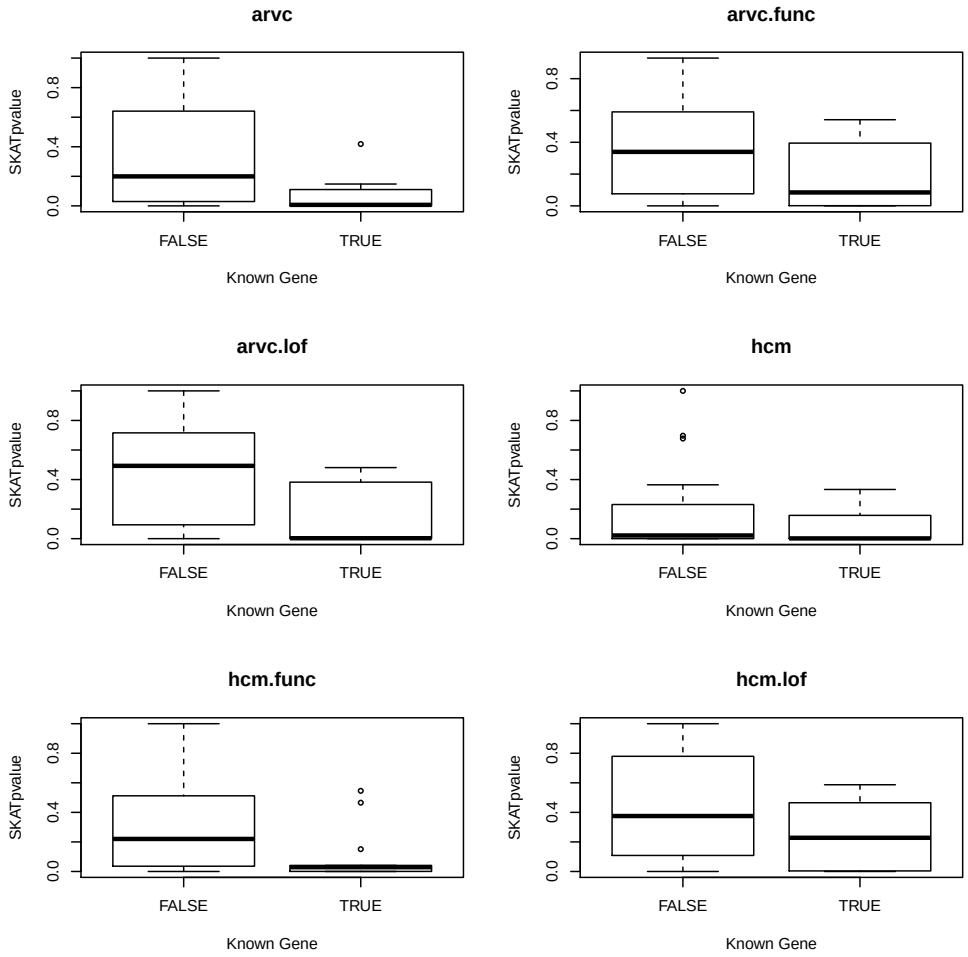


Figure 2.2: The predictive ability of known gene status. Boxplots showing pvalues for the SKAT tests on known and unknown candidate genes (True and False on the X-axis,respectively). Varying filters were used for determining variants included per gene test: no Functional Filter ('ARVC' and 'HCM' plots), Functional variants (non-synonymous and splicing). and Loss of Function - frameshift and stop-gain or stop loss (LOF)).

outcome, here phenotype. The idea that the effect size might increase as the MAF lowered was examined (Figure 2.4). The distribution of SNP effect sizes and Odds Ratios was also calculated (Figure 2.5) .

This section aimed to gain improve our understanding of ARVC and HCM. While it was not possible to identify a clear difference in the pattern of SNP effect sizes or Odds Ratios between these conditions, the trend of increasing effect with decreasing MAF was clear for both.

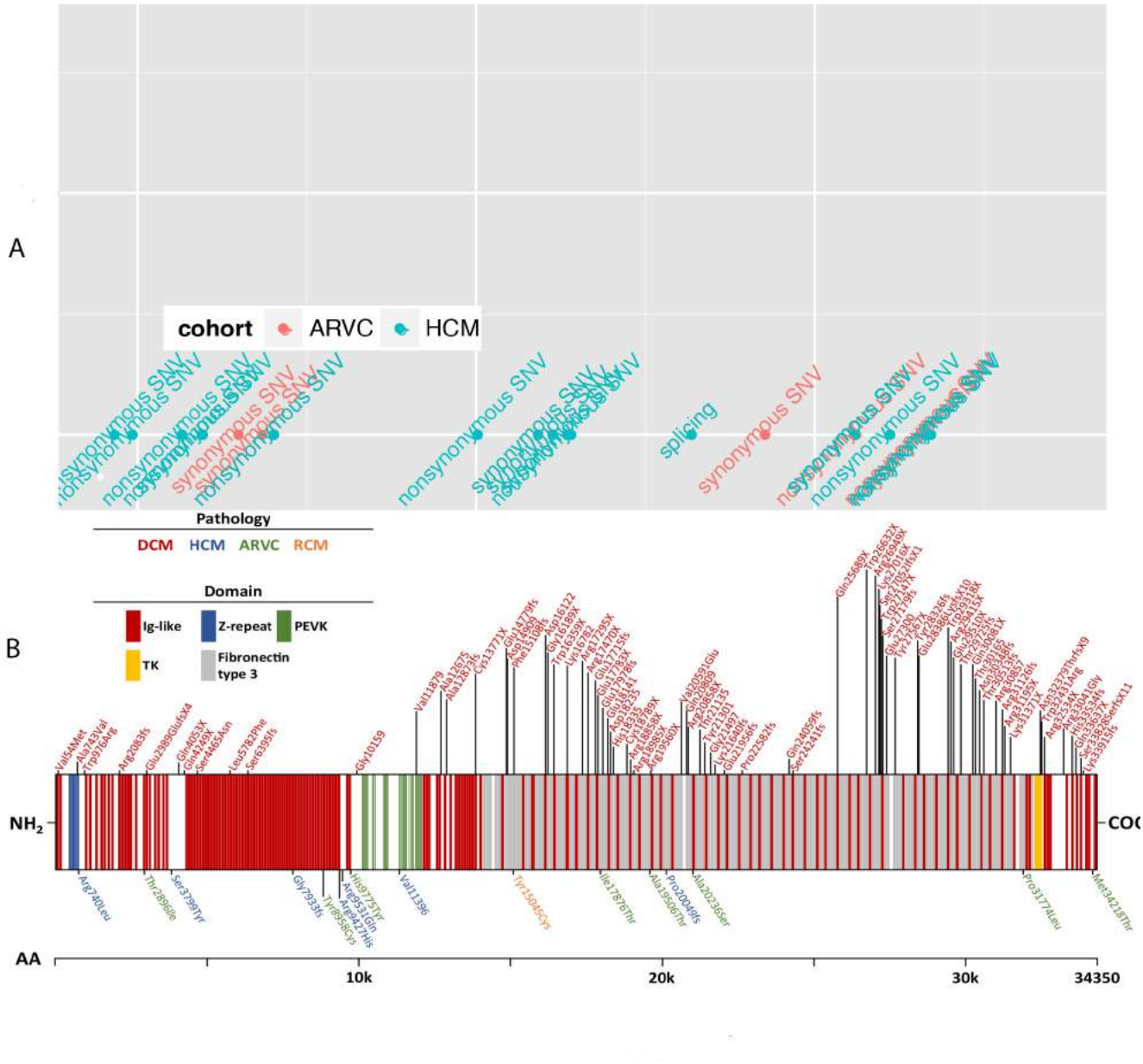


Figure 2.3: The distribution of candidate variants across the Titin coding sequence. (A) Variants with a pvalue of $\leq 1 * 10^{-4}$ from the case control analysis were plotted in relation to their respective locations across the *TTN* gene, where the X-axis is the position of the variant along the Titin sequence and the y axis is unused. The color of each mutation represents the associated pathology. (B) Reproduced from [Neiva-Sousa et al., 2015]. Mutations associated with cardiomyopathies distributed along the canonical *TTN* sequence (UniProtKB: Q8WZ42-1). The type of domain in *TTN* is represented by the color of each block in the sequence. Abbreviations: DCM is dilated cardiomyopathy,HCM hypertrophic cardiomyopathy, ARVC arrhythmogenic right ventricular cardiomyopathy, RCM restrictive cardiomyopathy, Ig immunoglobulin, PEVK region rich in proline (P), glutamate (E), valine (V) and lysine (K), TK titin Ser/Thr kinase

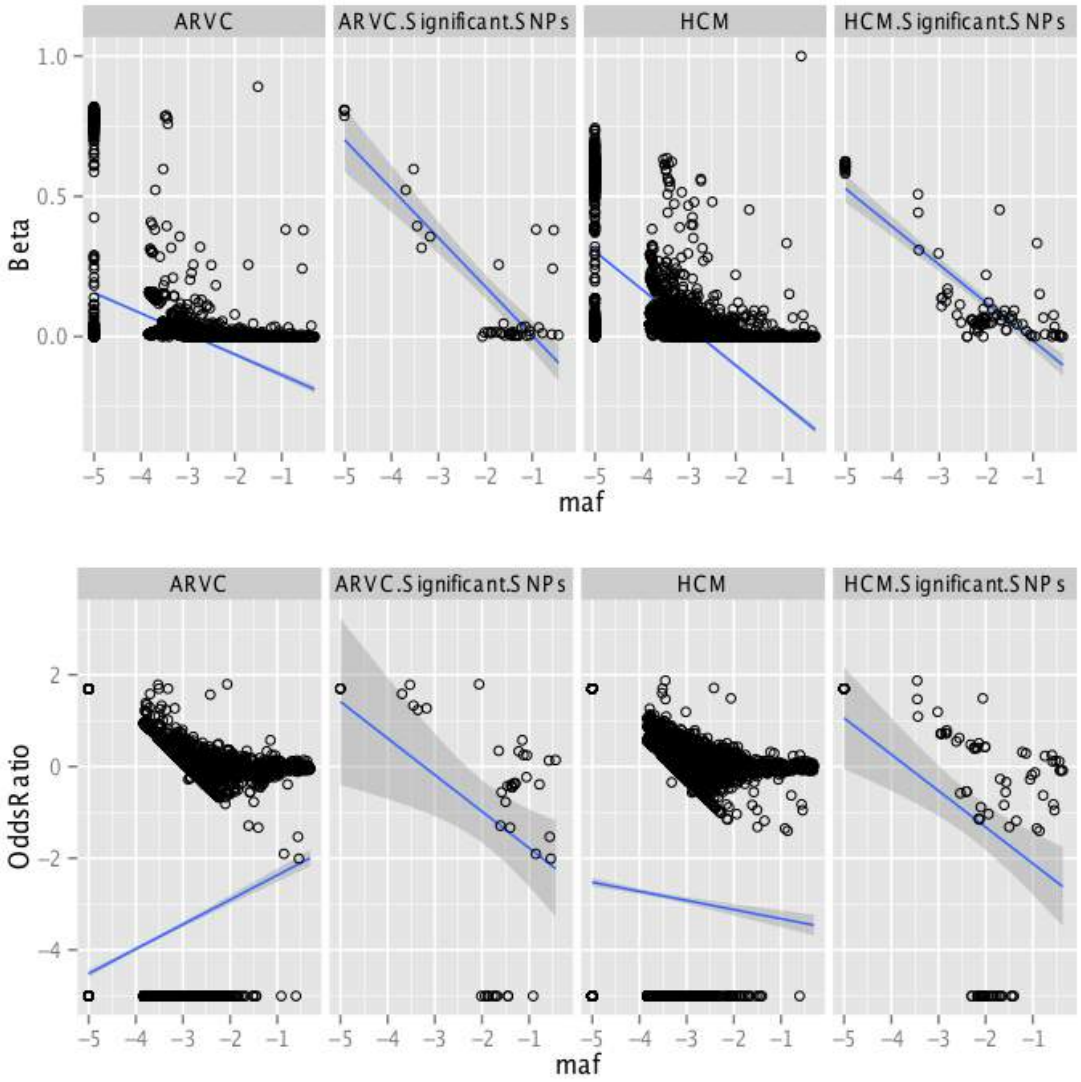


Figure 2.4: Characterising variants in ARVC and HCM across the Minor Allele Frequency (MAF) spectrum. 'ARVC' shows all variants while 'Significant' indicates those with a pvalue of ≤ 0.0001. Top Panel - X axis: log10 of control MAF, Y axis: Squared regression coefficients of the model phenotype SNP. Loess regression and standard error shown as line with grey perimeter. Lower Panel: Y axis represents the log10 of the risk Odds Ratio.

	gene	funcSKAT	funcSKATO	funcFisher	funcChiSq	LoFSKAT	LoFSKATO	LoFFisher	LoFChiSq
1	*PKP2*	1.82E-33	2.73E-43	1.75E-30	3.84E-44	8.70E-34	5.26E-53	5.40E-34	5.35E-48
2	*DSG2*	1.47E-10	4.57E-13	3.29E-09	2.14E-11	6.45E-02	1.59E-02	2.11E-02	2.39E-02
3	*TMEM43*	4.53E-04	7.46E-04	3.47E-02	5.23E-02				
4	*DSP*	1.42E-03	8.24E-04	1.01E-02	1.03E-02	4.76E-09	8.05E-16	3.38E-09	1.29E-12
5	*TPM1*	1.28E-03	1.19E-03	2.36E-03	1.23E-03				
6	*TTN*	2.32E-03	5.12E-03	7.54E-01	7.92E-01	4.56E-01	6.43E-01	6.89E-01	6.90E-01
7	*LMNA*	3.95E-03	7.58E-03	8.63E-02	1.58E-01				
8	*KCNE2*	9.81E-03	1.36E-02	5.72E-02	8.89E-02				
9	*ACTC1*	1.91E-02	1.91E-02	1.57E-01	3.45E-01				
10	*PNN*	1.85E-02	3.26E-02	2.63E-01	3.67E-01				
11	*DSC2*	4.04E-02	6.73E-02	6.07E-01	6.80E-01	1.64E-01	2.19E-01	5.02E-01	6.15E-01
12	*TTN-AS1*	6.57E-02	9.82E-02	3.23E-01	7.87E-01				
13	*SCN5A*	1.21E-01	1.35E-01	1.31E-01	1.47E-01				
14	*CAV3*	1.15E-01	1.81E-01	4.09E-01	9.79E-01				
15	*JUP*	2.29E-01	2.07E-01	1.78E-01	2.53E-01				
16	*DES*	2.34E-01	2.18E-01	1.35E-01	2.67E-01				
17	*RBM20*	1.79E-01	3.06E-01	8.11E-01	9.57E-01				
18	*MYBPC3*	1.95E-01	3.36E-01	4.44E-01	6.12E-01	6.26E-01	6.26E-01	1.00E+00	1.00E+00
19	*KCNE1*	3.89E-01	3.81E-01	2.78E-01	4.20E-01				
20	*GJA1*	2.51E-01	3.82E-01	1.00E+00	1.00E+00				
21	*KCNQ1*	2.60E-01	3.92E-01	3.08E-01	5.24E-01				
22	*MYL2*	3.98E-01	3.98E-01	4.03E-01	9.67E-01				
23	*LDB3*	2.96E-01	4.77E-01	8.02E-01	1.00E+00				
24	*CSRP3*	8.94E-01	4.89E-01	1.00E+00	7.24E-01				
25	*MYH6*	3.55E-01	5.56E-01	6.62E-01	8.38E-01	8.26E-01	6.71E-01	1.00E+00	1.00E+00
26	*KCNJ2*	3.83E-01	5.65E-01	1.00E+00	1.00E+00				
27	*PLEC*	7.99E-01	5.94E-01	4.93E-01	4.65E-01				
28	*TNNI3*	8.01E-01	6.39E-01	1.00E+00	1.00E+00				
29	*TCAP*	8.79E-01	6.47E-01	7.60E-01	6.46E-01				
30	*PLN*	8.19E-01	6.61E-01	1.00E+00	1.00E+00				
31	*KCNH2*	8.26E-01	6.71E-01	1.00E+00	1.00E+00				
32	*TGFB3*	6.52E-01	7.24E-01	1.00E+00	8.23E-01				
33	*CASQ2*	9.73E-01	7.37E-01	8.57E-01	7.82E-01				
34	*PDLIM3*	5.24E-01	7.39E-01	1.00E+00	9.36E-01				
35	*TNNT2*	6.29E-01	8.27E-01	1.00E+00	1.00E+00				
36	*ANK2*	6.70E-01	8.53E-01	7.82E-01	7.47E-01				
37	*VCL*	6.94E-01	8.81E-01	1.00E+00	9.33E-01				
38	*MYH7*	7.20E-01	9.07E-01	1.00E+00	8.92E-01	8.95E-01	5.95E-01	1.00E+00	1.00E+00
39	*RYR2*	7.19E-01	9.13E-01	8.87E-01	9.10E-01	7.88E-01	6.23E-01	1.00E+00	1.00E+00
40	*PKP4*	7.71E-01	1.00E+00	8.45E-01	1.00E+00	1.06E-02	1.39E-02	7.39E-02	1.21E-01
41	*MYL3*	8.08E-01	1.00E+00	1.00E+00	1.00E+00				

Table 2.4: ARVC Gene based Results. Here, each gene has multiple pvalues as the variants included were varied as was the exact statistical test used. 'func' refers to variants that were predicted to have any impact on the transcribed DNA sequence, including synonymous, non-synymous and splicing changes. 'LoF' refers to Loss of Function which is frameshift, stopgain, stoploss or conserved splicing. Both of these variant sets for each gene was then tested with the Fisher, SKAT and SKATO tests. Absent values indicate no variants remain after filtering.

	gene	funcSKAT	funcSKATO	funcFisher	funcChiSq	LoFSKAT	LoFSKATO	LoFFisher	LoFChiSq
1	*MYBPC3*	3.73E-23	1.23E-47	2.12E-31	2.07E-33	9.79E-07	1.07E-15	1.95E-14	4.34E-14
2	*MYH7*	8.30E-22	8.00E-41	2.32E-25	1.52E-27	4.54E-01	6.30E-01	1.00E+00	1.00E+00
3	*TNNI3*	4.21E-06	5.31E-09	4.44E-08	4.11E-08				
4	*TTN*	1.50E-04	4.12E-04	1.77E-01	1.86E-01	5.89E-01	7.25E-01	3.67E-01	3.60E-01
5	*TPM1*	4.19E-02	6.08E-04	2.28E-04	2.50E-04				
6	*CSRP3*	1.37E-03	2.30E-03	2.89E-02	3.27E-02				
7	*ACTC1*	5.37E-02	4.40E-03	8.46E-03	1.29E-02				
8	*TNNT2*	6.21E-03	7.44E-03	2.81E-02	3.48E-02				
9	*PKP2*	1.56E-02	7.55E-03	7.59E-03	7.74E-03	1.43E-02	4.24E-04	8.85E-04	1.31E-03
10	*DSP*	7.21E-03	1.44E-02	1.00E+00	9.93E-01	4.72E-01	4.72E-01	1.00E+00	1.00E+00
11	*PDLIM3*	1.07E-02	2.12E-02	1.40E-01	1.70E-01				
12	*MYL2*	1.01E-01	2.87E-02	3.10E-02	5.24E-02				
13	*ANK2*	3.27E-02	6.36E-02	9.22E-01	1.00E+00				
14	*RBM20*	3.97E-02	7.28E-02	1.96E-01	2.04E-01				
15	*SCN5A*	4.28E-01	8.47E-02	4.84E-02	5.91E-02				
16	*PLN*	1.23E-01	9.24E-02	7.86E-02	1.48E-01				
17	*TTN-AS1*	1.08E-01	1.04E-01	1.13E-01	2.20E-01				
18	*VCL*	1.13E-01	1.16E-01	1.01E-01	1.12E-01				
19	*CASQ2*	5.40E-01	1.59E-01	1.37E-01	1.62E-01				
20	*MYL3*	1.26E-01	2.12E-01	1.00E+00	1.00E+00				
21	*DSG2*	1.26E-01	2.22E-01	5.14E-01	6.08E-01	5.03E-01	4.24E-01	5.56E-01	5.42E-01
22	*RYR2*	3.59E-01	2.42E-01	1.58E-01	1.68E-01	2.62E-01	3.97E-01	6.06E-01	8.68E-01
23	*TCAP*	1.78E-01	2.86E-01	1.00E+00	1.00E+00				
24	*DSC2*	2.88E-01	3.54E-01	1.43E-01	1.71E-01	3.57E-01	2.91E-01	1.20E-02	3.54E-02
25	*MYH6*	4.14E-01	3.66E-01	1.84E-01	2.16E-01	2.48E-01	3.77E-01	5.91E-01	7.70E-01
26	*TMEM43*	5.56E-01	3.84E-01	3.68E-01	3.78E-01				
27	*KCNH2*	2.53E-01	3.85E-01	5.86E-01	7.26E-01				
28	*KCNE1*	4.24E-01	3.85E-01	3.68E-01	3.81E-01				
29	*GJA1*	2.86E-01	4.12E-01	4.45E-01	5.65E-01				
30	*LDB3*	2.55E-01	4.18E-01	3.80E-01	4.32E-01				
31	*KCNE2*	3.80E-01	4.71E-01	3.82E-01	5.73E-01				
32	*LMNA*	4.27E-01	5.02E-01	5.29E-01	4.83E-01				
33	*PLEC*	3.06E-01	5.06E-01	9.17E-01	9.91E-01				
34	*PKP4*	3.36E-01	5.40E-01	4.12E-01	5.15E-01	2.78E-01	3.79E-01	5.36E-01	1.00E+00
35	*TGFB3*	3.74E-01	5.70E-01	5.92E-01	7.59E-01				
36	*KCNJ2*	6.12E-01	6.46E-01	6.74E-01	7.58E-01				
37	*DES*	6.72E-01	7.23E-01	5.50E-01	7.29E-01				
38	*KCNQ1*	5.33E-01	7.26E-01	1.00E+00	1.00E+00				
39	*JUP*	5.55E-01	7.67E-01	1.00E+00	1.00E+00				
40	*CAV3*	6.56E-01	7.82E-01	1.00E+00	1.00E+00				
41	*PNN*	7.32E-01	8.98E-01	7.87E-01	1.00E+00				

Table 2.5: HCM Gene based Results. Here, each gene has multiple pvalues as the variants included were varied as was the exact statistical test used. 'func' refers to variants that were predicted to have any impact on the transcribed DNA sequence, including synonymous, non-synymous and splicing changes. 'LoF' refers to Loss of Function which is frameshift, stopgain, stoploss or conserved splicing. Both of these variant sets for each gene was then tested with the Fisher, SKAT and SKATO tests. Absent values indicate no variants remain after filtering.

	Gene	Nb.variants	PercentOfTotal
1	*MYBPC3*	64	61
2	*TTN*	22	21
3	*MYH7*	8	7
4	*TPM1*	5	4
5	*MYH6*	2	1
6	*PLN*	2	1
7	*TNNT2*	1	1

Table 2.6: Number of LOF variants in HCM Candidate Genes

	Gene	Nb.variants	PercentOfTotal
1	*PKP2*	28	37
2	*DSP*	14	18
3	*DSC2*	11	14
4	*TTN*	9	12
5	*DSG2*	4	5
6	*PKP4*	3	4
7	*JUP*	2	2
8	*LMNA*	2	2
9	*DES*	1	1

Table 2.7: Number of LOF variants in ARVC Candidate Genes

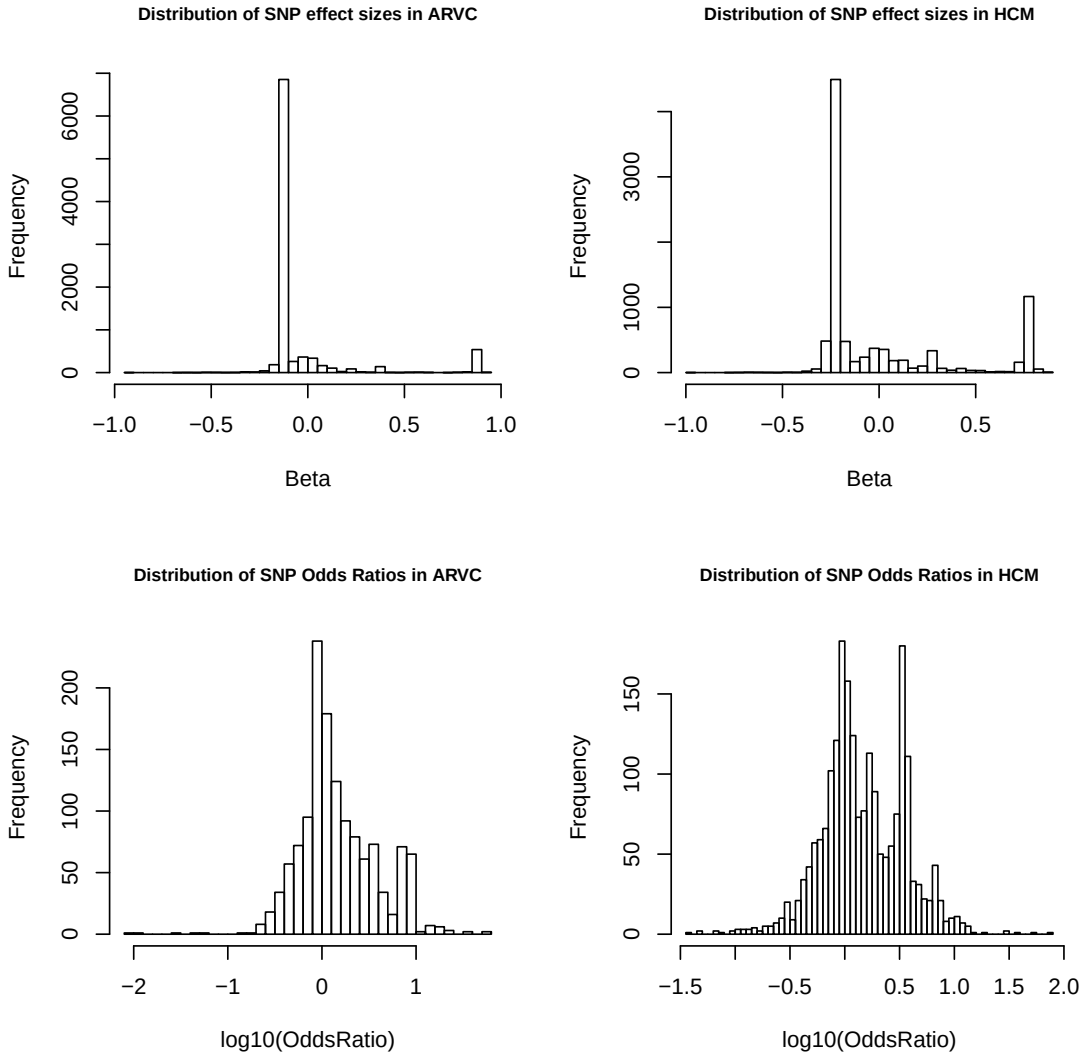


Figure 2.5: The distribution of effect sizes and odds ratios in ARVC and HCM.

2.3 Discussion

2.3.1 Molecular autopsy of Sudden Cardiac Death patients

The molecular autopsy of the SCD cohort showed that for a small portion of people who die suddenly with no identifiable cause, even post-mortem, rare potentially pathogenic DNA variants harboured in genes associated with SCD may be the answer. Throughout this study, rare variants were thought more likely candidates because common variants with large enough effects to cause SCD were thought quite unlikely to exist. Reasonable disease associated ion channel mutations were found in 3 (5%) probands. 2 of these families had private mutations; mutations that are found in only that respective family. These were R1623Q and V411M and were concomitant with negative clinical screens. The third family had a *RyR2* mutation with a malignant history but no clear phenotype. A further 6 (10%) had rare ion channel variants which have previously been associated with Brugada Syndrome and Long QT syndrome. They were not extremely rare in controls however, at $0.02 - 0.5\%$. Finally, eight (14%) had rare or very rare cardiomyopathy variants of unknown significance.

The case control tests of these samples against UCL-ex were suggestive of an excess of rare variants in cases, but not statistically significant. This might be due to the fact that our sample size was relatively limited. A large number of families refused access to DNA and were therefore not included in the study. 24% of the DNA that was collected was unsuitable for analysis. This highlights the need for improved consistent guidelines as to how to adequately store tissue samples for further investigations, potentially years later. Traditional methods of sample storage, such as Formalin Fixed Paraffin Embedded (FFPE), have been shown to have inadequate DNA preservation [Tournier et al., 2012]. Alternatively, it may be because our knowledge of the relevant genes and the impact different variants will have is still incomplete. This is made more difficult because the hearts of SCD probands are structurally normal and typically there is no associated ante-natal clinical phenotype data.

That being said, for the families that did receive a diagnosis, even if post mortem, some solace can be gained. It might result in increased participation with proactive family genetic screening, altered reproductive

choices or simply relief from the resolution of the mystery as to the cause of death. In conclusion, while the approach of molecular autopsy is undoubtedly useful in SCD, its limited diagnostic yield means it can augment but is not a viable replacement of traditional clinical testing.

2.3.2 ARVC and HCM case control analysis.

407 ARVC and 955 HCM samples were sequenced for the genes listed in Table 4 in the Appendix. These genes are known or thought to be associated with either HCM or ARVC. Single variant and gene based case control tests were then performed against 3587 UCL-ex ethnicity and phenotype matched controls.

Table 2.2 lists the top associations for the targeted sequencing ARVC analysis (Fisher pvalue of $\leq 1 * 10^{-4}$) . This is dominated by *PKP2* and *DSG2*, which is in agreement with the literature. The most significantly associated variant is the previously reported splice site altering rs193922674 SNP [Gerull et al., 2004]. The role of Titin in ARVC however is less clear. A recent study on 38 ARVC families identified 8 unique *TTN* variants across 7 families. One of these variants, Thr2896Ile, perfectly segregated the ARVC phenotype in a large family [Taylor et al., 2011]. This group has gone further and associated that *TTN* variant carriers are at greater risk of supraventricular arrhythmias and conduction disease [Brun et al., 2014]. While intriguing, this has yet to be independently verified so more work needs to be done.

Table 2.3 lists the variants most associated with HCM. As is the case with ARVC, these data largely agree with the currently understood genetic architecture of HCM; one largely driven by *MYBPC3*. A manual examination of the rs1805123 *KCNH2* variant showed that in multiple samples, this multiallelic locus is low RD. This reduces our confidence in this being a true call, but the concordance between our control MAF (0.25) and that of ExAC (0.19) in comparison to the HCM MAF of 0.38 does make this a candidate worthy of following up. At the very least, it highlights the importance of a stringent QC process controlling for as many parameters as possible.

RBM20 was first associated with Dilated Cardiomyopathy in 2009 [Brauch et al., 2009]. Since then, it has further been found that it is a splicing regulator of *TTN* [Li et al., 2010]. In December 2015, the first human induced Pluripotent Stem Cell (hIPSC)model of *RBM20* model was published [Wyles et al., 2015].

This took dermal fibroblasts from two patients carrying the R636S missense variant and transformed them into hIPSC derived cardiomyocytes. These cell lines exhibited a downregulation of *RBM20* concomitant with a downregulation of the adult isoform of *TTN* N2B and upregulation of the foetal *TTN* N2BA isoform. *LDB3*, *CAMK2D* and *CACNA1C* genes were also affected. The net result of these changes was that the sarcomeres, when developed, exhibited increased sarcomeric length and decreased width. This makes *RBM20*'s role in HCM a relatively plausible one. In our data, rs35141404 has a HCM MAF of 0.27 and has a frequency in our ARVC cohort of 0.17 and our control frequency is 0.18. The ExAC MAF is 0.15, agreeing with our control frequency. However, the low call rate we observed for this variant is cause for concern; a concern somewhat lessened by the fact that the ExAC data does report the same issue at this locus. Thus, this variant is an interesting candidate and will be further assessed with Sanger sequencing and when our sample size increases.

Figure 2.4 characterises the SNP effect size and the odds ratio across the MAF spectrum for both ARVC and HCM. Here, the effect size is the coefficient from the logistic regression when phenotype is modelled as the outcome against each SNP. When all SNPs are examined together, no clear pattern is visible. However, when one restricts this to those with a significant pvalue, it becomes clear that both the risk odds ratio and the effect size has an inverse relationship with the MAF in controls.

The single variant tests were accompanied with gene based tests. Functional variants, non-synonymous, frameshift or stop site altering and those with a MAF ≤ 0.05 were retained. Testing was then performed in a number of different ways: basic Fisher & χ^2 tests that counted all unfiltered variants and SKAT and SKATO. SKAT aggregates individual SNP test statistics in a given set (here, a gene) and then calculates the corresponding pvalue. *PKP2* was the most associated ARVC gene with a SKAT p-value of $1.82 * 10^{-33}$, followed by *DSG2* ($1.47 * 10^{-10}$) [Table 2.4]. These were the only statistically significant genes found here. Furthermore, the top HCM genes were found to be *MYBPC3* ($3.73 * 10^{-23}$) and *MYH7* ($8.3 * 10^{-22}$) [Table 2.5]. The SKAT-O Loss of Function (LOF) test differs from SKAT for ARVC in that it indicates that *DSP* plays a more important role than *DSG2*. *DSP* based ARVC can follow an autosomal dominant or autosomal recessive model of inheritance and may be associated with palmoplantar keratadoma and Carvajal disease

[Sen-Chowdhry et al., 2007]. Having access to such patient phenotype data would be informative as it would enable us to create a more refined picture of the architecture of the subtypes of ARVC.

Chapter 3

Analysis of Copy Number Variants in Hypertrophic Cardiomyopathy

3.1 Introduction

Professor Perry Elliott, Dr. Luis Lopes and Dr. Petros Syrris at the Heart Hospital, University College London, have collected a cohort of 505 patients with HCM [Lopes et al., 2013b]. A targeted panel of 41 genes (Table 5) was chosen based on the knowledge, at the time the array was designed, of the genetic basis of HCM and ARVC. The average Read Depth across the 2.1Mb region was 120. A variant was included in the filtered list of potentially disease-causing variants if it was both rare (defined as having a MAF of $\leq 0.5\%$) and non-synonymous, LOF or a splice site variant. Excluding Titin, 152 candidate variants were identified, 89 of which were novel.

The role of copy-number variants (CNV) as a cause of hypertrophic cardiomyopathy (HCM) is poorly studied. The aim of this chapter was to use high-throughput sequence (HTS) data combined with a read-depth strategy, to screen for CNVs in cardiomyopathy-associated genes in a large consecutive cohort of HCM patients. Identified CNVs were then validated by Array Comparative Genome Hybridisation (aCGH) A large portion of this chapter is published in Lopes et al. [2015]. I did all of the CNV analysis: the read

depth approach with ExomeDepth, designed the probes for the aCGH and employed the SVD approach of CoNIFER.

3.2 Methods & Materials

3.2.1 Patients and Clinical Evaluation

The study cohort was comprised of 505 patients diagnosed with HCM at the Heart Hospital, University College London, UK. A 12-lead Electrocardiogram (ECG), echocardiography and exercise testing were used in the diagnosis. A left ventricular wall thickness on two-dimensional ECG of $\geq$ 13mm, after correction for age, sex and size was the diagnostic threshold used.

3.2.2 Targeted gene enrichment and high-throughput sequencing

In total, the 41 target genes spanned a 2.1Mb region of genomic DNA [Lopes et al., 2013b] per patient. This included exonic, intronic and certain regulatory regions, 20 of which were either associated with HCM, ARVC or Dilated Cardiomyopathy (DCM), a related phenotype. The remaining genes are implicated in other cardiomyopathies or arrhythmias. The capture and sequencing methodology used for the first 233 patients has been reported in detail previously [Lopes et al., 2013b]. From the 234th patient onwards, successive updated versions of the Agilent sample preparation protocol were used according to the manufacturer's instructions. The main changes referred to smaller initial quantities of genomic DNA (200 ng to 3 mg), use of Agilent enzymes and reagents throughout the protocol, optimisations of hybridisation steps and replacement of in-solution PCR procedure with an on-bead PCR method. Introduction of additional SureSelect indexes allowed multiplexing of 16 samples in a single pool. The resulting index-tagged sample pools were sequenced on the Illumina HiSeq 2000 system. Cluster generation on Illumina cBot was carried out according to the manufacturer's protocol. A total of 128 HCM samples (16 multiplexed samples * 8 lanes) were sequenced (100 bp, paired end) per instrument run, using standard methods (Illumina).

The paired-end reads were then aligned using the Novalign Software V.2.7.19 against the hg19 human

reference genome. Once duplicated regions were excluded with Picard MarkDuplicate Tool, indels and SNPs were called with SAMtools [Li et al., 2009]. A minimum genotype quality threshold of Phred score 30 was implemented to curate the resultant variant list. A Phred value of 30 is equivalent to a 99.9% [Ewing et al., 1998] base call accuracy rate. Annovar was used for sample annotation [Wang et al., 2010].

3.2.3 ExomeDepth

It is recommended practice to compare a sample against another sample or set of samples to estimate a normalised measure of RD [Plagnol et al., 2012]. This is more accurate than creating an intra-individual measure as there is a high degree of exon to exon variability. By comparing the target region/sample against this null, one can calculate the likelihood of the presence of a Duplication or Deletion. ExomeDepth, the R based implementation of a RD approach, fits a beta binomial model that builds an optimised reference set that maximises the CNV detection power [Plagnol et al., 2012]. This can work on even small (1 - 2 exons) CNVs even in the midst of technical variability.

The samples were sequenced in 22 batches. To minimise the effect of the resultant technical variability on the CNV calling, and to generate sample sets of a size that maximises the CNV calling algorithm, the samples were analyzed by these batches (Figure 3.1). A script was written in R that did this using the ExomeDepth R package (on CRAN) (Figure 3.1). Sorted, indexed BAM files that had duplicate reads removed with PICARD MarkDuplicate were used. For each sample, all other samples in its set were used as potential controls. ExomeDepth then identifies the optimum set of sample(s) from this group to compare the test sample against. This is done by identifying reference samples that are comparable to the test sample. RD similarity is the main criterion.

3.2.4 CoNIFER

Another widely used bioinformatic approach to call CNVs from targeted sequencing is CoNIFER [Krumm et al., 2012]. CoNIFER takes as input sample Reads per kilobase per milllion (RPKM) values. This is a sequence length standardised measure of the number of reads per region. CoNIFER then uses Singular Value

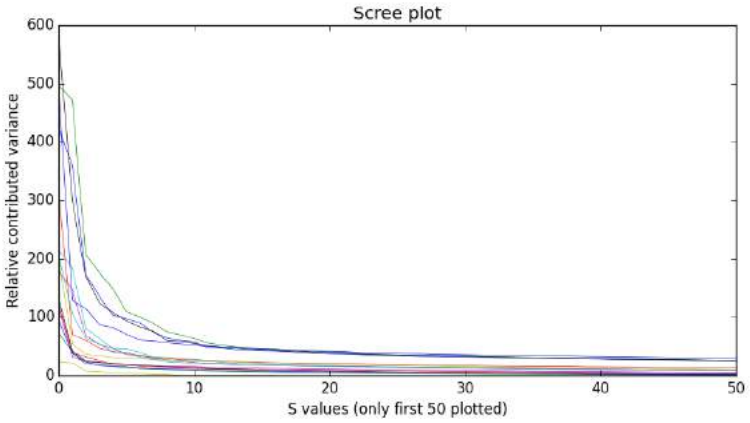
Figure 3.1: QR Codes.(A) The HCM sample sequencing plate information. https://github.com/CianMurphy/Upgrade/blob/master/bamFileList.csv (B) The ExomeDepth R script used to generate the CNV calls with ExomeDepth. https://github.com/CianMurphy/Upgrade/blob/master/ExomeDepth.R

Decomposition (SVD) to remove biases in the data. If X is the mean and standard deviation standardised RPKM values in the form of an exon by sample matrix, then the SVD of X takes the form $X = USV^T$. SVD is related to PCA in that the singular values S are the square roots of the eigenvectors of the covariance matrix XX^T. One can visualise the proportion of variance explained by each of the components (Singular Values) as a screeplot (Figure 3.2). Typically, K components are removed based on the inflection point of the scree plot to eliminate as much as noise as possible. For the screeplot included here, a K of 4 was thus chosen. CoNIFER can detect CNVs of 3 exons or larger. These data were then exported to R, where the DNAcopy package was used to implement the Circular Binary Segmentation (CBS) algorithm. This is a more sensitive segmentation algorithm than the inbuilt one in CONIFER. CBS recursively splits chromosomes into either two or three subsegments based on a maximum t-statistic. A reference distribution is used to decide whether or not to split is estimated by permutation.

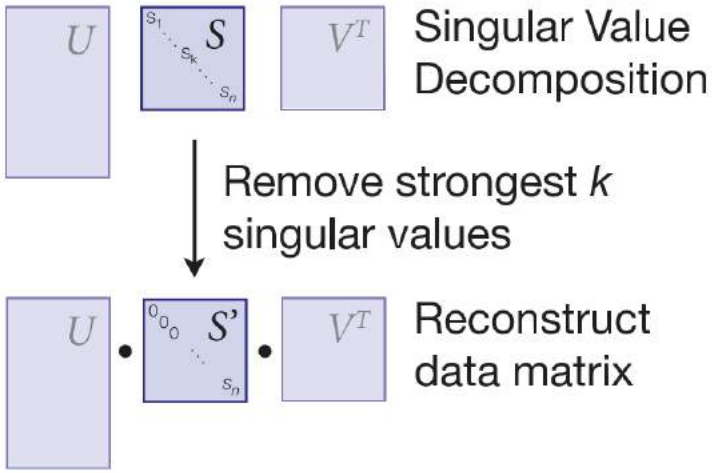
The different methods to detect CNVs have been rigourously compared against each other. In one such study, it was found that ExomeDepth had higher sensitivity than CONTRA, XHMM and CoNIFER [Tan et al., 2014]. Therefore, for a given set of samples, one would expect ExomeDepth to pick up the most CNVs, albeit with a higher false positive rate. This fact helped guide our experimental design here in that we used ExomeDepth as a first pass and combined CoNIFER and aCGH to subsequently validate the calls.

3.2.5 Array CGH

Comparative Genome Hybridisation (CGH) is a technique whereby you differentially fluorescently label two DNA samples [Oostlander et al., 2004]. They may come from different individuals, be a tumour pair or any



(a) Screeplot of SVD-RPKM values from a plate of 84 HCM samples.



(b) A graphical representation of the procedure for removing K SVDs.

Figure 3.2: CoNIFER analysis: Removing the components of the Singular Value Decomposition that disproportionately contribute to the variance.

other combination. Classically, the fluorescent dyes Cyanine-3 (Cy3) and Cyanine (Cy5) are used as their emission spectra are readily distinguishable. Both the test DNA and the reference DNA are then hybridized to cloned DNA fragments that have been spotted in a gridlike fashion on a glass slide, the "Array" portion of array CGH(aCGH). Subsequently, CNVs will be visible by a measurable difference in the emission spectra of the spots.

An aCGH was designed to validate the CNVs called by ExomeDepth R script and to verify that its algorithm identified all CNVs. This was done via the Agilent eArray server (https://earray.chem.agilent.com/earray/). This was designed to cover 2.1Mb of sequence across the target genes, with one probe every 100bp (Figure 3.3). Once the aCGH probe set was designed, it was submitted to Agilent with the 12 samples of interest. The array was built and once the samples were processed the data was sent back to us.

The data are in the form of probe intensity ratios (typically in the $\log_2$ scale). An experiment without measurement or normalization errors run on a normal CNV null clone would yield a Log_2 ratio of 0 because the test and reference sample would be equivalent [Guha, 2008]. The Log_2 ratio of a heterozygous deletion is $\text{Log}_2(1/2) = -1$ and a heterozygous gain is $\text{Log}_2(3/2) = 0.58$.

To determine if the aCGH called any CNVs in the 12 samples, the log2 ratios were processed in R. SnapCGH, aCGH and limma were the principal packages used. A brief overview of the process is as follows: Firstly, the data are read into R and a valid object is created to store it. The data are mined for the array positional information for the clones. In this aCGH experiment, Cy5 was used to fluoresce the reference sample, so this results in the addition of a design vector with a value of -1 (Cy3 for reference would be given +1). The next step is to control for the background intensity for each spot to improve the resolution later on. The "minimum" method in snapCGH was used which simply subtracts the background value from that of the foreground. The data are then normalized, before the segmentation model is fitted. This fits a homogenous Hidden Markov Model (HMM). Segmentation is vital as it splits the data into probe sets that share the same DNA copy number [Ben-Yaacov and Eldar, 2008]. Segmentation has a tendency to fit states that have similar means, which can obfuscate the true copy number state of the sample. One method to ameliorate this is to merge states that have means within a defined threshold. Once this has completed, the

data are ready for plotting as identifying CNVs works well from a visualization approach.

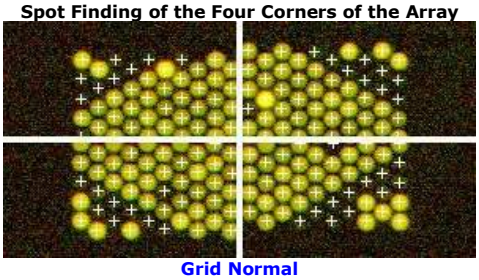
Figure 3.3: QR Code for HCM CNV aCGH validation probes.*https://github.com/CianMurphy/Upgrade/blob/master/CGH_ca*

The 2 colour aCGH was performed by Agilent. Their proprietary Feature Extraction software creates detailed quality control reports, in addition to the probe intensity ratios. The Spot Finding image in Figure 3.4 allows you to determine whether or not the spots have been correctly located centrally on the array. If this was not the case, the results would be unreliable. The bottom left table in this figure details population attributes. If this showed a greater than expected number of non-uniform or population outliers then this would indicate a hybridization/wash step error. The plots above this table show the spatial distribution of both the population and non-uniform outliers on the array. This is a useful method to determine if a given subset of samples are outliers. The panel 'Evaluation Metrics for CGH_QCMT_Sep09' describes array attributes such as background noise and the signal to noise ratio and offers suggestions as to which do or do not meet their quality thresholds. This is a guide to assist further evaluation. Here, it is noted that the red background noise is high, but in practise there was no issue in calling CNVs from this sample. The histogram of Signals Plots the number of points in discrete intensity bins against the log2 of the processed signal to give the shape and level of signal distribution. Figure 3.5 top plot shows the spatial distribution of the positive and negative log ratios. A lack of discernible pattern in this is what is expected. Figure 3.5 bottom plot shows the log of the red background corrected signal against the log of the green background corrected signal for non-control inlier features. The linearity or curvature of this is a guide for choosing the appropriate background method choices as this plot should be linear. The intersection of the red horizontal and vertical lines shows the position of the median signal while the numbers below the plot indicate the number of non control features that have a background corrected signal less than zero. Overall, these reports show the aCGH was performed to a high quality, allowing confidence in their data.

QC Report - Agilent Technologies : 2 Color CGH

Date	Wednesday, August 14, 2013 - 17:24	Sample(red/green)	
User Name	scan	FE Version	10.7.3.1
Image	0938_UCL_255024210023_S01 [1_1]	BG Method	Detrend on (NegC)
Protocol	CGH_107_Sep09 (Read Only)	Multiplicative Detrend	True
Grid	050242_D_F_20130617	Dye Norm	Linear
Saturation Value	65526 (r), 65525 (g)		

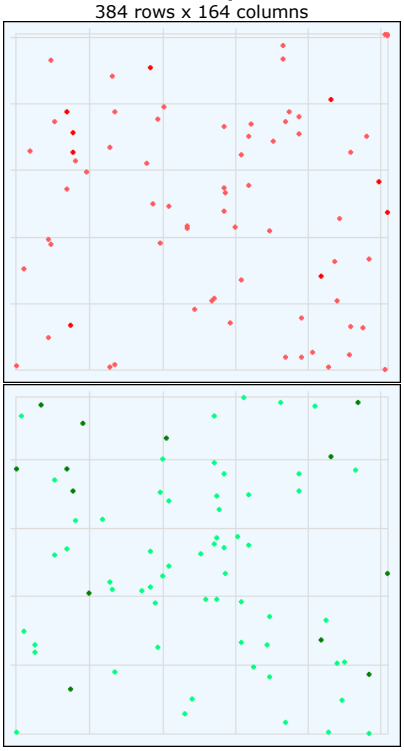
Spot Finding of the Four Corners of the Array



Grid Normal

Outlier Numbers with Spatial Distribution

384 rows x 164 columns



Red FeaturePopulation Red Feature NonUniform

Green FeaturePopulation Green Feature NonUniform

Feature	Red	Green	Any	% Outlier
Non Uniform	9	13	16	0.03
Population	64	59	108	0.17

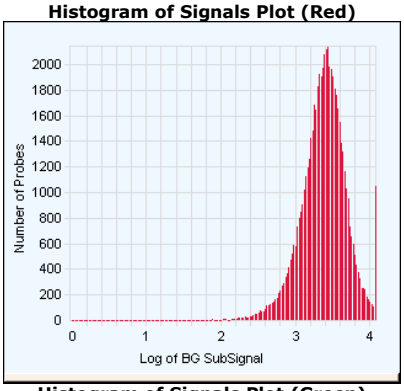
Evaluation Metrics for CGH_QCMT_Sep09 :

Excellent (7) ; Good (3) ; Evaluate (1)

Metric Name	Value	Excellent	Good	Evaluate
IsGoodGrid	1.00	>1	NA	<1
AnyColorPrcntFeatNonUn...	0.03	<1	1 to 5	>5
DerivativeLR_Spread	0.12	<0.20	0.20 to 0.30	>0.30
gRepro	0.10	0 to 0.05	0.05 to 0.20	<0 or >0.20
g_BGNoise	8.02	<5	5 to 10	>10
g_Signal2Noise	168.67	>100	30 to 100	<30
g_SignalIntensity	1352.55	>150	50 to 150	<50
rRepro	0.09	0 to 0.05	0.05 to 0.20	<0 or >0.20
r_BGNoise	16.20	<5	5 to 10	>10
r_Signal2Noise	128.10	>100	30 to 100	<30
r_SignalIntensity	2074.67	>150	50 to 150	<50

♦ Excellent ♦ Good ♦ Evaluate

Histogram of Signals Plot (Red)



Histogram of Signals Plot (Green)

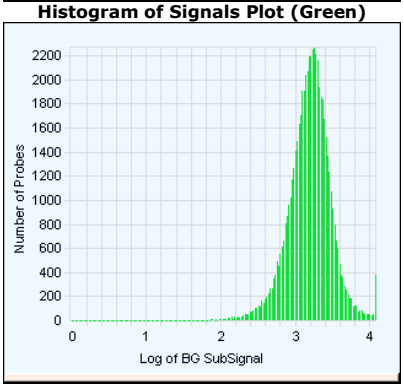


Figure 3.4: Quality metrics for sample 0938_UCL_255024210023_S01_CGH_107_Sep09 from the Agilent aCGH.

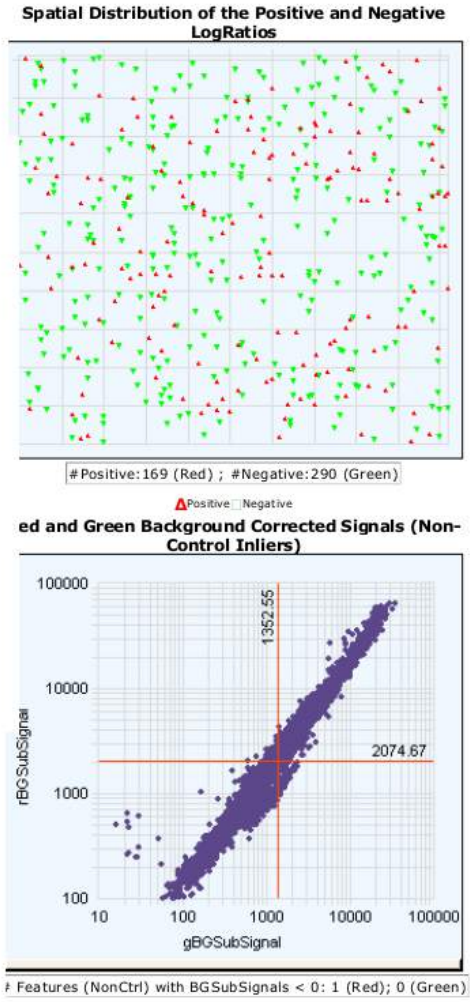


Figure 3.5: Quality metrics for sample 0938_UCL_255024210023_S01_CGH_107_Sep09 from the Agilent aCGH.

3.3 Results

3.3.1 ExomeDepth HCM CNVs

ExomeDepth was the first method used to identify CNVs in this cohort [Plagnol et al., 2012]. In brief, the mean value of the per-sample average RD in the exonic target region across the samples was 348.09±142.59. Combining all samples and taking the mean value across all samples, 92.41% of the target region was covered to a RD of 15 or more. A 2010 study generated a 42 million probe tiled microarray that identified 11,700 CNVs, thought to include 80-90% of common CNVs [Conrad et al., 2010]. These data are incorporated into ExomeDepth and are used as an initial filter to remove common variants on the basis that they are unlikely to be disease causing. At our selected confidence threshold level after filtering, 12 CNVs in 12 patients (2.4% of the 505 cohort) were identified using ExomeDepth.

3.3.2 aCGH Validation of the HCM CNVs

The Log_2 ratios have been normalised, segmented and quantified. To plot the varying intensities from probe to probe, a region file was first drawn up demarcating the location of the 41 genes of interest. 4 of the 12 most likely CNVs from ExomeDepth, in 4 patients (0.8% of the cohort) were validated by aCGH:

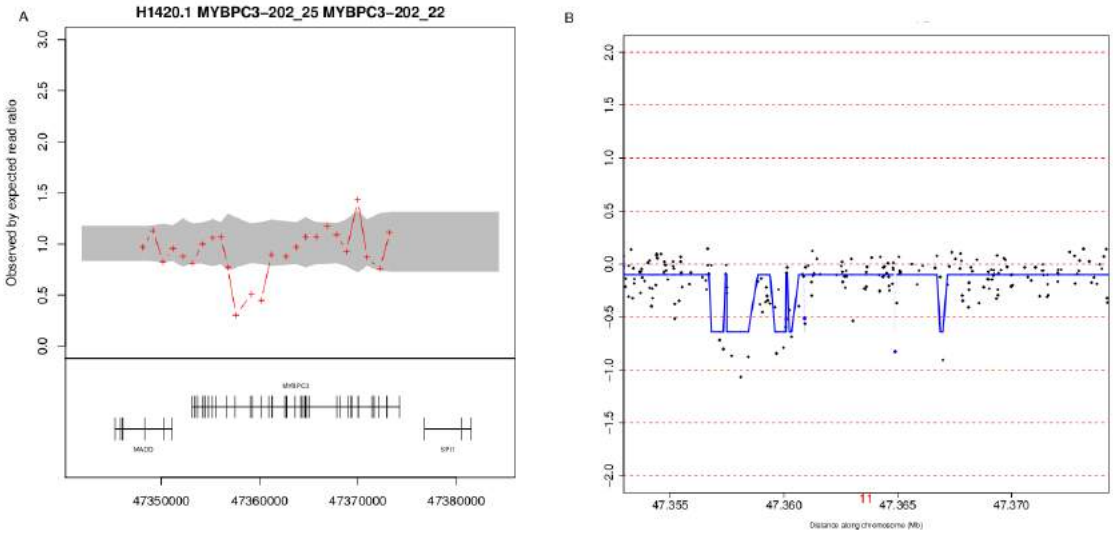
- one large deletion in *MYBPC3* (involving 4 exons) shown in Figure 3.6
- one duplication of the entire *TNNT2* gene shown in Figure 3.7
- one large deletion in *PDLIM3* (involving the first 4 exons) shown in Figure 3.8
- and one large duplication in *LMNA* (involving 5 exons) in Figure 3.9

Three of them did not harbour any variant in a potentially causal sarcomere gene and one is a carrier of a variant of unknown significance in *TNNT2*.

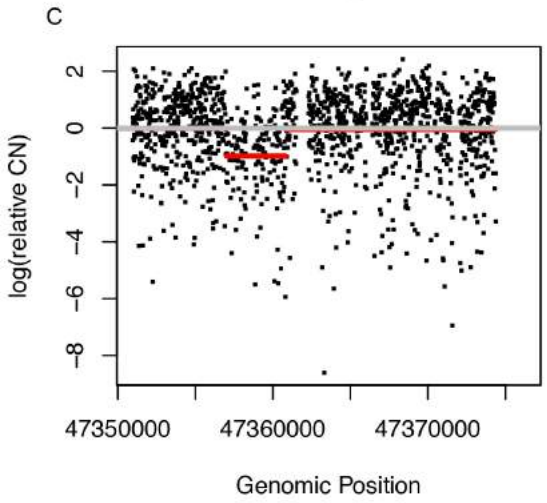
Eight CNVs were not validated by the aCGH analysis, including three single exon duplications and one single exon deletion in *MYBPC3*, two two-exon deletions and one single exon duplication in *TNNI3* and

one single exon duplication in *ACTC1* (Figures 3.10 - 3.17). Owing to the high probe density of the aCGH in these genes, my interpretation is that these 8 CNV calls are false positives. Nevertheless, I cannot exclude that some of these CNV calls are real but too small to be validated by other techniques. The aCGH did not identify additional CNV calls in these 12 samples.

Because the accuracy of CNV calling algorithms are limited [Tan et al., 2014] I compared the ExomeDepth CNV calls with the output of CoNIFER [Krumm et al., 2012]. Using the suggested settings, CoNIFER identified a much larger number of CNV calls (120 calls overall). CoNIFER called the 4 CNVs validated by the aCGH experiment but did not call any of the 8 CNVs not validated by the aCGH experiment (Figures 3.10 - 3.17). Owing to the intuitively excessive number of CNV calls, combined with the fact that a visual analysis of CoNIFER output plot was largely unconvincing, I assumed that owing to technical factors specific to this experiment, the false positive rate of CoNIFER was high and did not follow-up these calls.

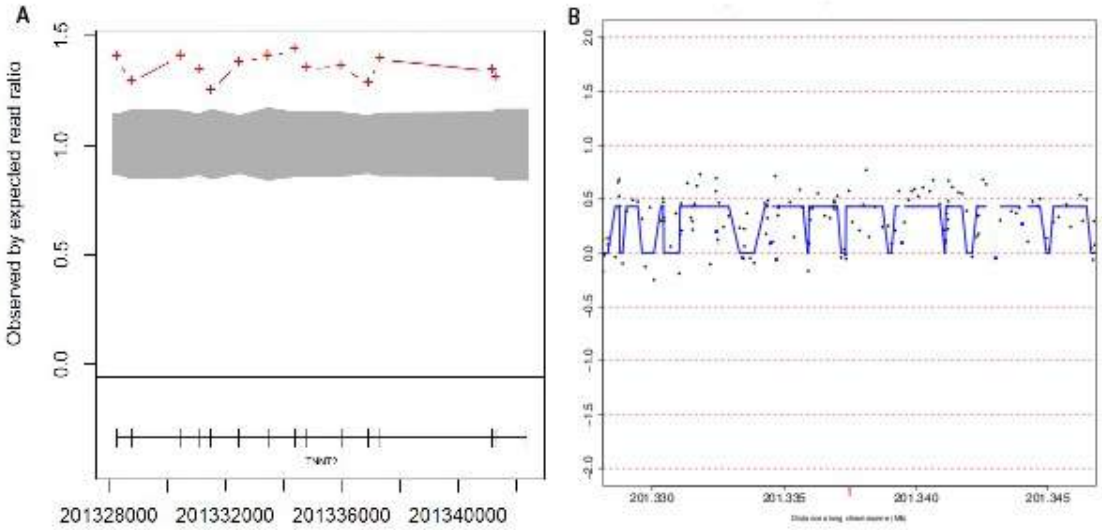


(a) Panel (A): ExomeDepth plot. The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons.(B) aCGH plot. The blue line represents the fitting of a homogenous Hidden Markov Model for Segmentation by snapCGH.

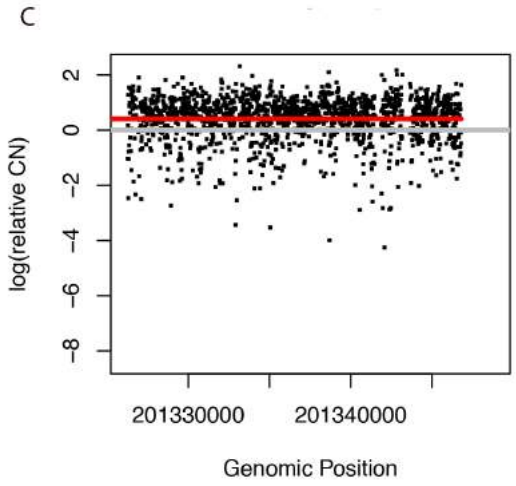


(b) Panel(C): The CNV as called by CoNIFER using Singular Value Decomposition. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

Figure 3.6: Deletion in *MYBPC3* in patient H1.

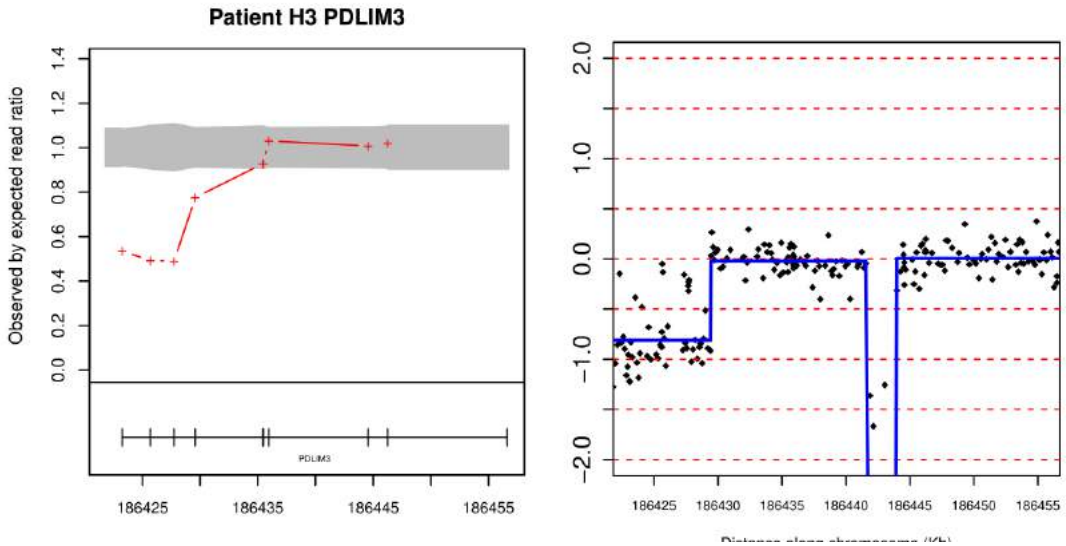


(a) Panel (A): ExomeDepth plot. The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) aCGH plot. The blue line represents the fitting of a homogenous Hidden Markov Model for Segmentation.

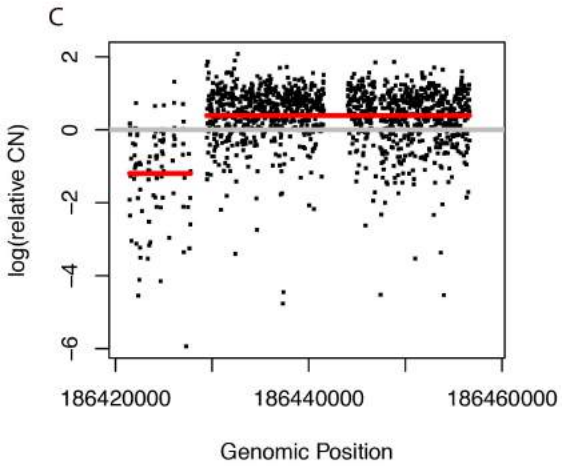


(b) Panel(C): The CNV as called by CoNIFER using Singular Value Decomposition. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

Figure 3.7: Patient H2 Exonic Duplication in *TNNT2*.

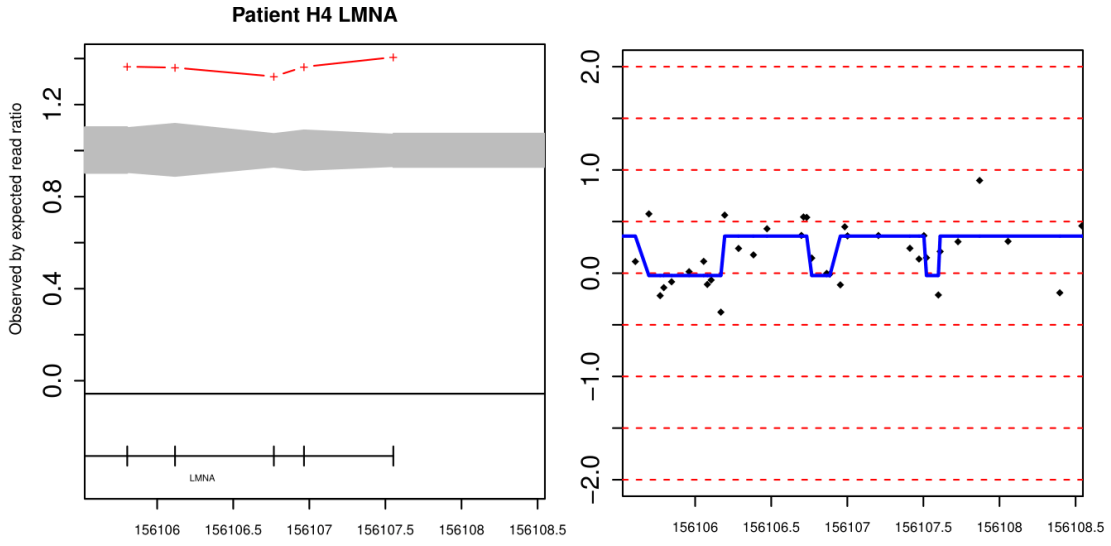


(a) Panel (A): ExomeDepth plot. The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) aCGH plot. The blue line represents the fitting of a homogenous Hidden Markov Model for Segmentation.

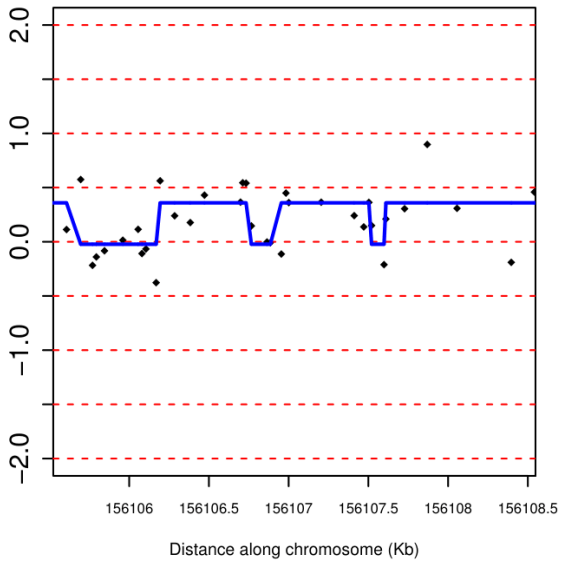


(b) Panel(C): The CNV as called by CoNIFER using Singular Value Decomposition. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

Figure 3.8: Patient H3 Exonic Duplication in *PDLIM3*.



(a) Panel (A): ExomeDepth plot. The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) aCGH plot. The blue line represents the fitting of a homogenous Hidden Markov Model for Segmentation.



(b) Panel(C): The CNV as called by CoNIFER using Singular Value Decomposition. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

Figure 3.9: Patient H4 Exonic Duplication in *LMNA*.

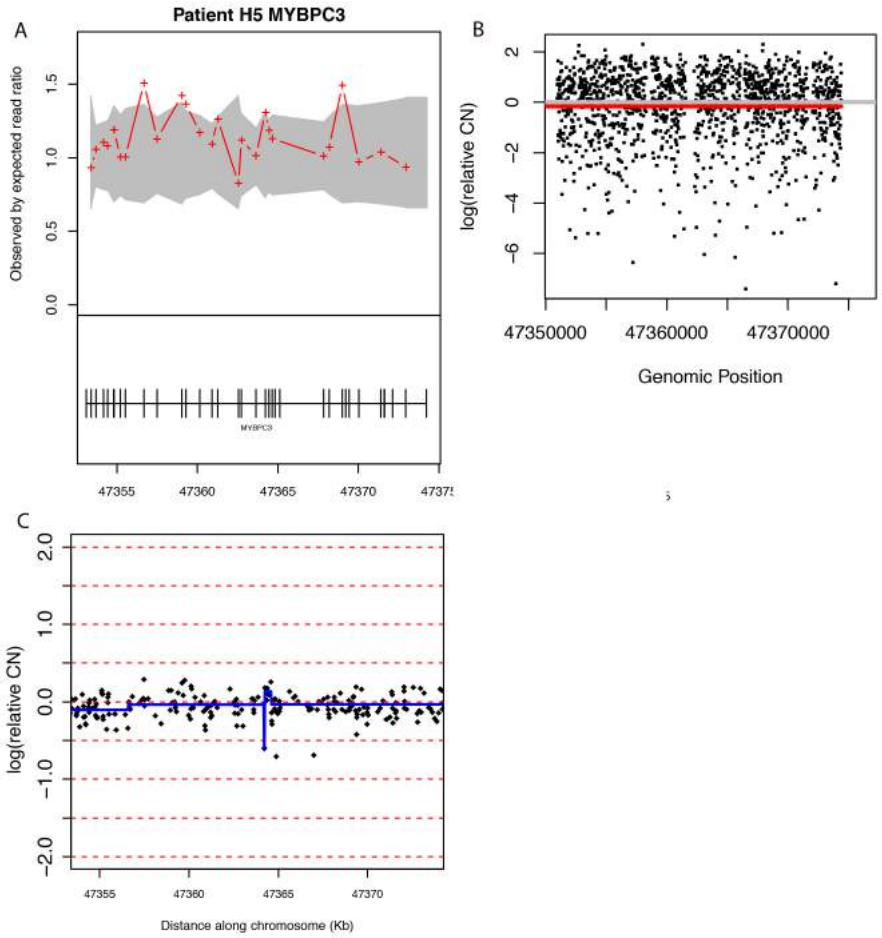


Figure 3.10: The first of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *MYBPC3* of patient H5 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

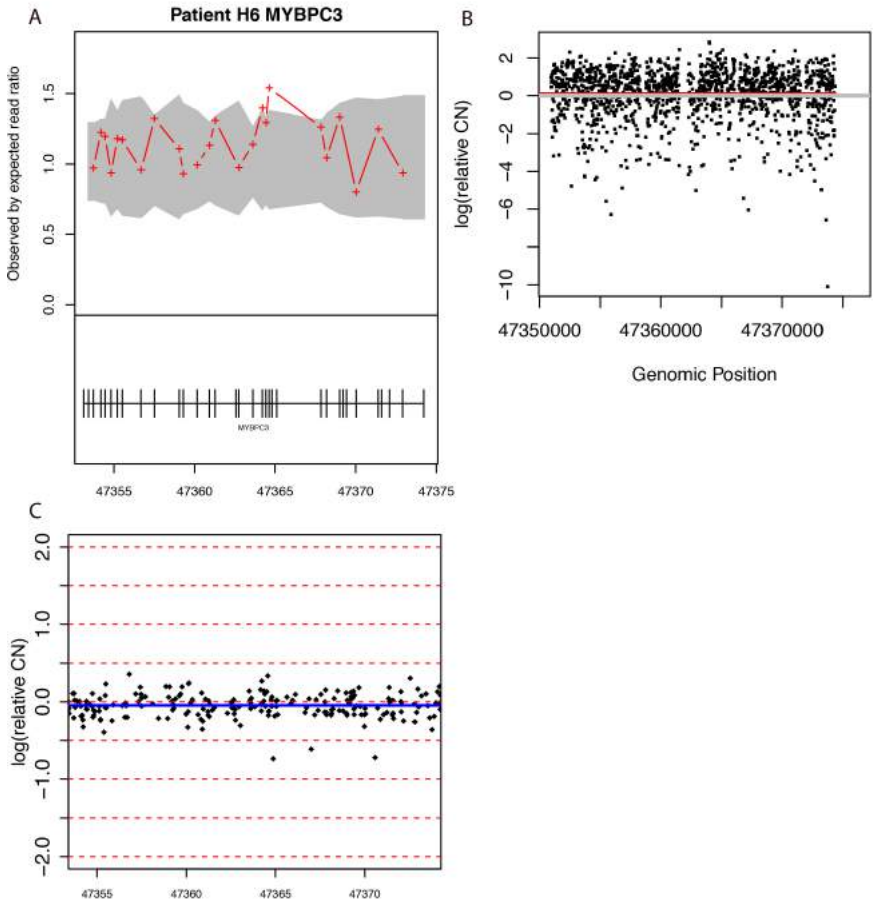


Figure 3.11: The second of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *MYBPC3* of patient H6 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

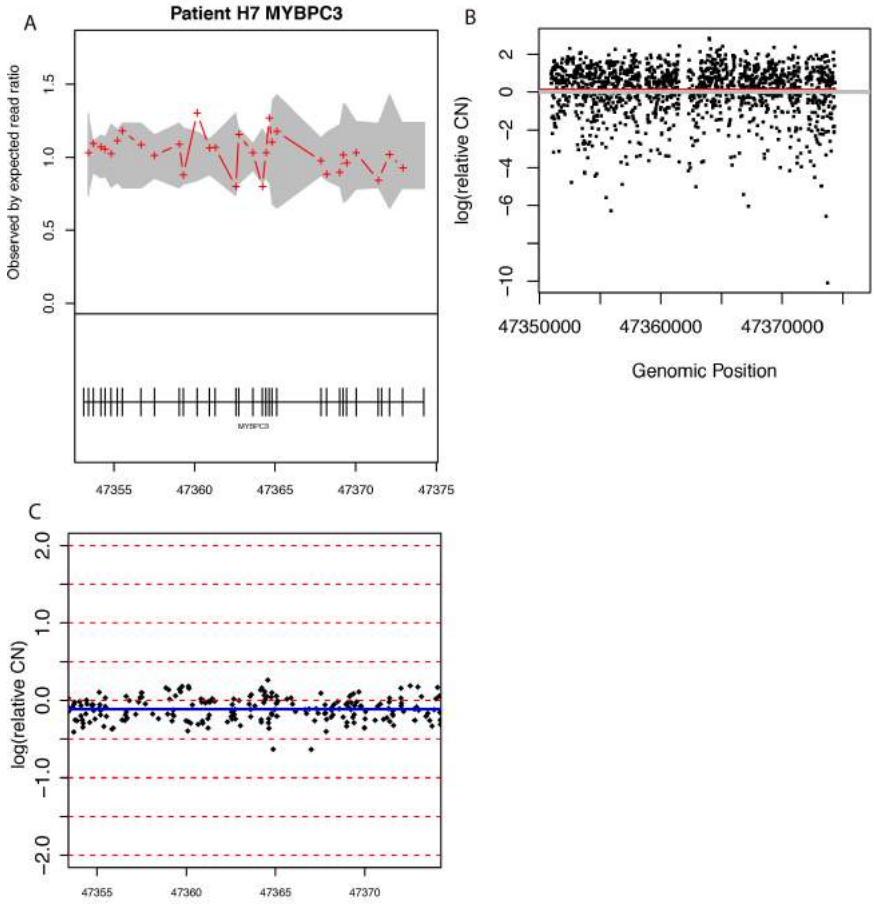


Figure 3.12: The third of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *MYBPC3* of patient H7 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

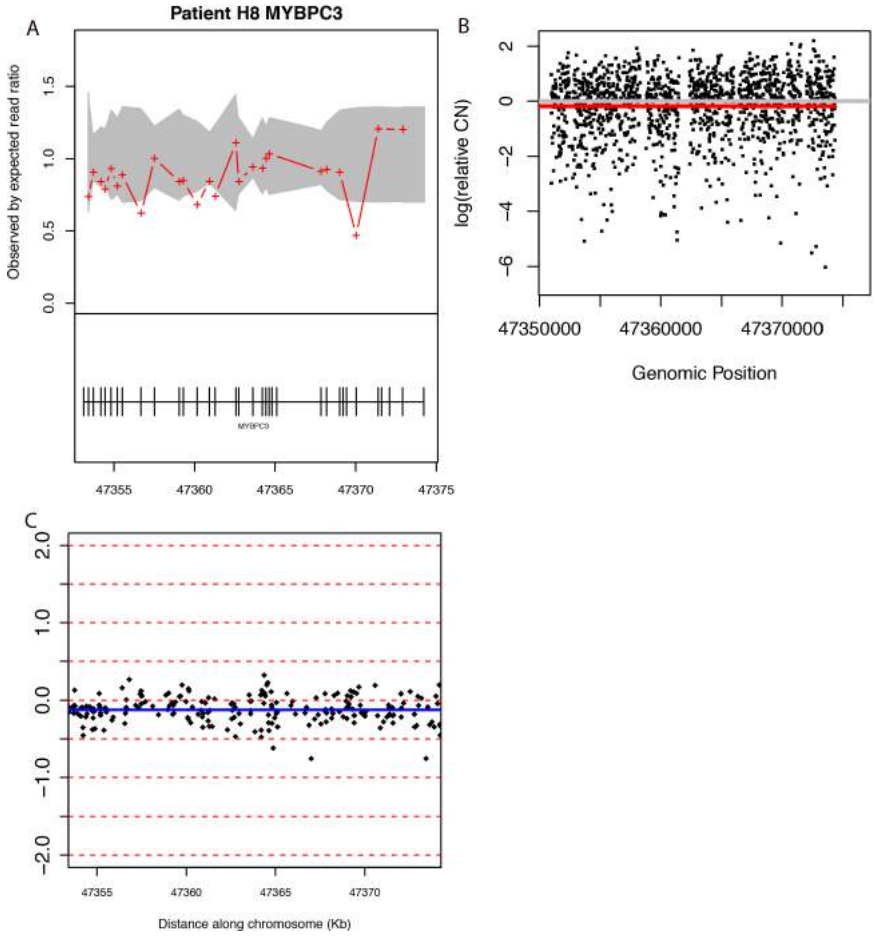


Figure 3.13: The fourth of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *MYBPC3* of patient H8 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

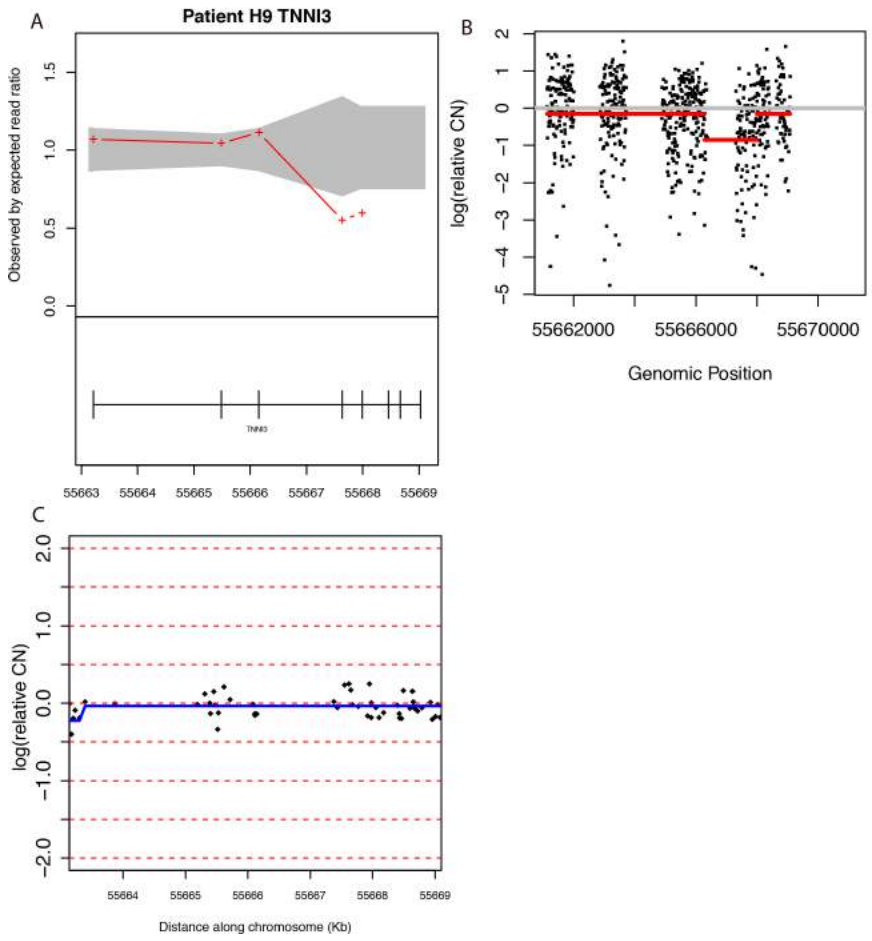


Figure 3.14: The fifth of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *TNNI3* of patient H9 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

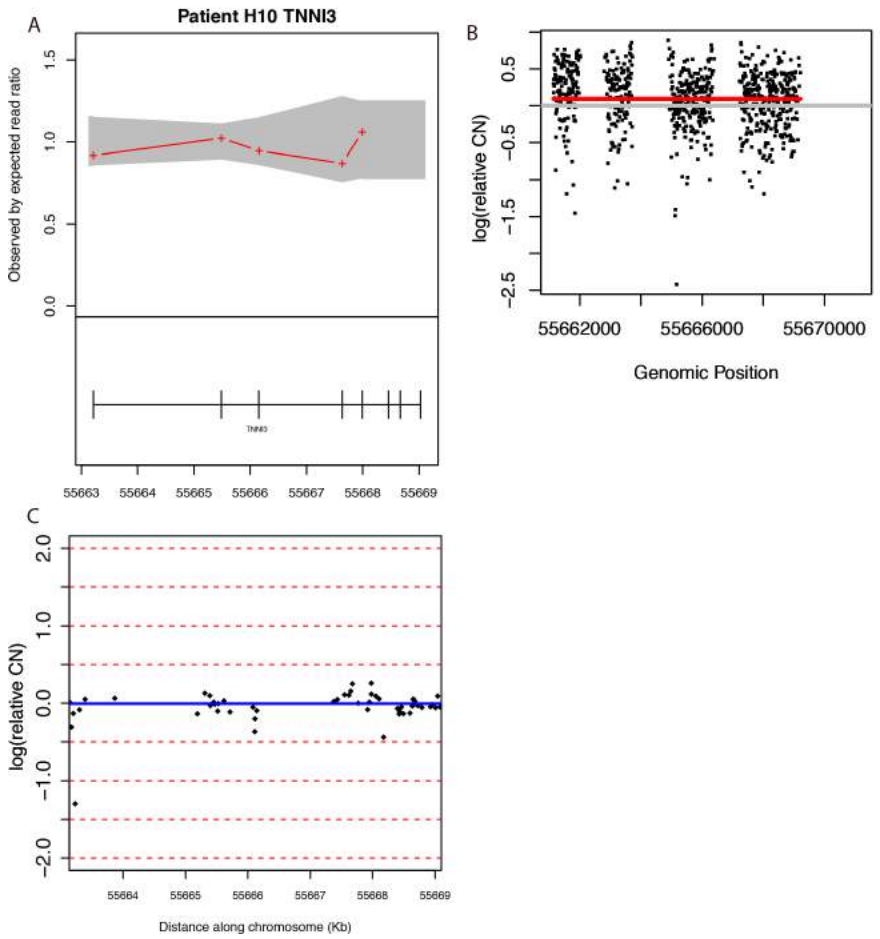


Figure 3.15: The sixth of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *TNNI3* of patient H10 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

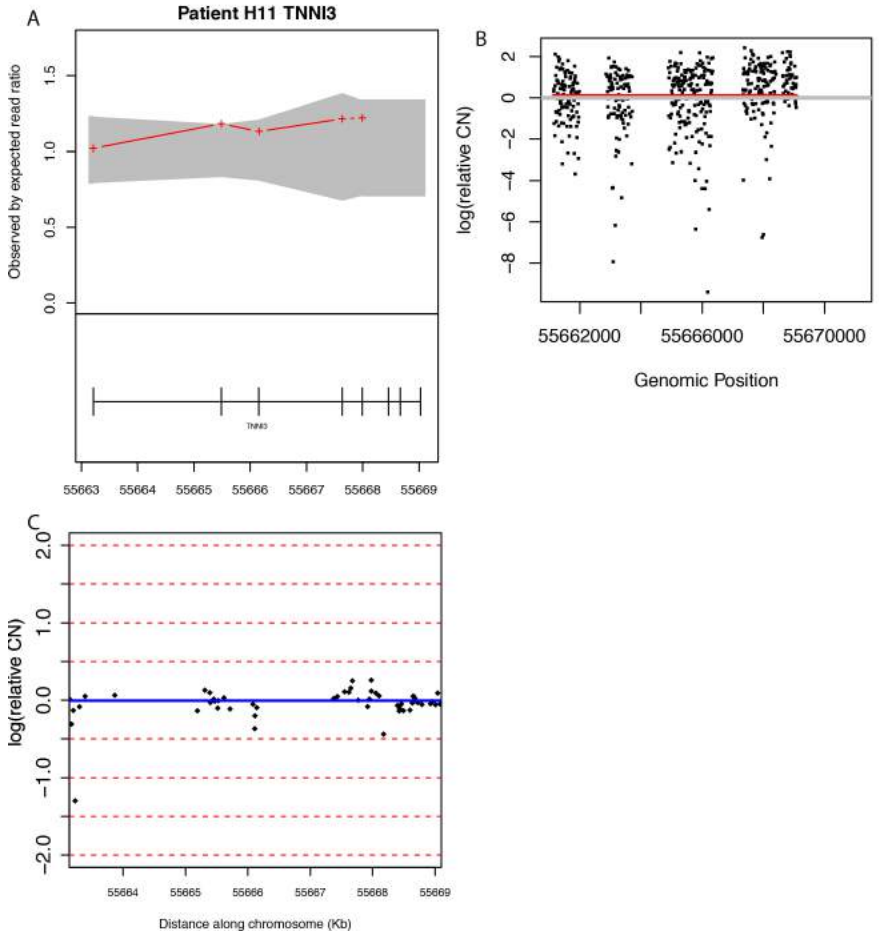


Figure 3.16: The seventh of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *TNNI3* of patient H11 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

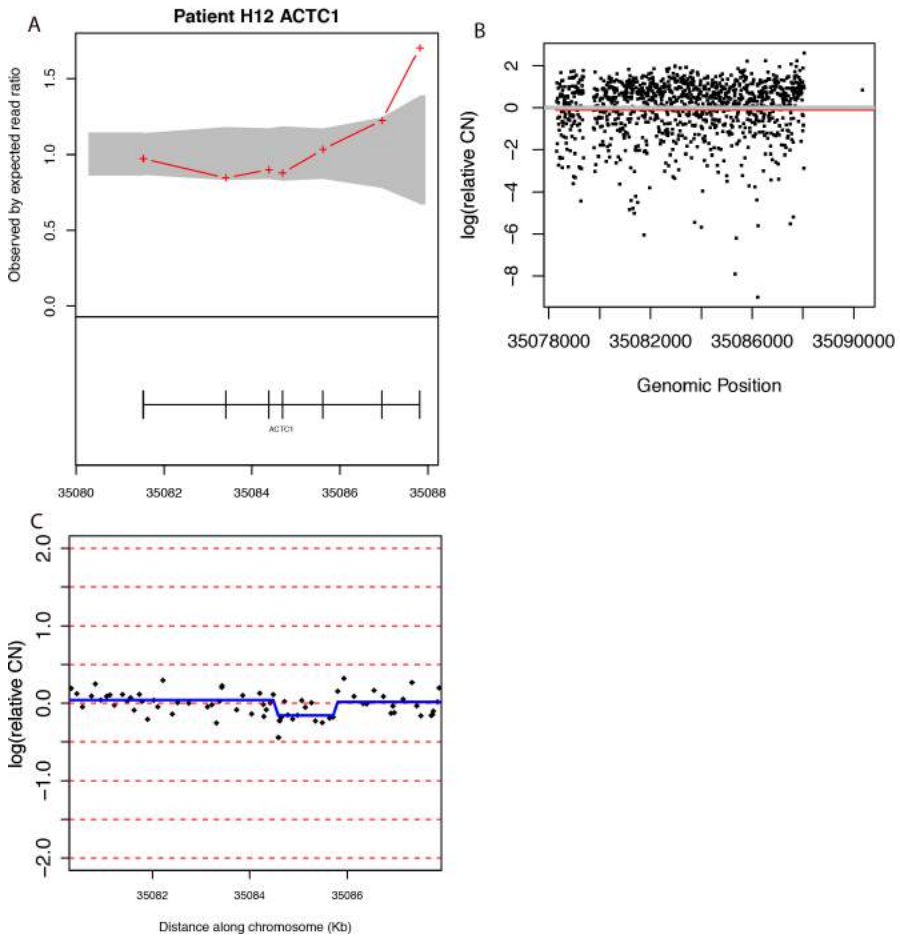


Figure 3.17: The eighth of eight unconfirmed CNVs called by ExomeDepth. Here, it called a duplication in the gene *ACTC1* of patient H12 (panel A). The red crosses indicate the ratio of observed/expected number of reads (RR). The grey-shaded area is the estimation of the 99% Confidence Interval for the RR in the absence of a CNV call. The X axis shows the affected gene plotted underneath, with the vertical lines showing the location of the exons. (B) The subsequent aCGH/snapCGH (Panel B) plot. The blue horizontal line represents the fitting of a homogenous Hidden Markov Model for Segmentation. (C) CoNIFER plot. The red line is the result from the Circular Binary Segmentation algorithm applied to the CoNIFER data.

3.4 Discussion

505 consecutive and unrelated patients that have been diagnosed with HCM underwent targeted exome sequencing of 41 genes that are either known or thought to be involved with disease pathogenesis. In addition to the SNP analysis discussed in [Lopes et al., 2013b], a RD based CNV identification strategy was used here. This was motivated by evidence that approximately 50-60% of HCM patients remain genetically undiagnosed [Lopes et al., 2013a]. This methodology was however hindered by the recognized difficulty in short-read approaches that negatively affects sensitivity/specificity [Duan et al., 2013].

In an attempt to deal with these technical issues, two other approaches were used to validate the 12 calls made by ExomeDepth. Both CoNIFER, which utilised an SVD-RPKM approach, and the Array CGH cytogenetic method validated 4/12 calls. This is in line with the previously high false positive rate of all available algorithms [Duan et al., 2013; Tan et al., 2014]. Using this multi-step design, we detected and validated potentially disease causing CNVs in 0.8% of samples. This has direct implications for diagnostic and counselling services: some patients without mutations found through direct sequencing may still have transmissible CNVs in sarcomeric protein genes.

Information about the contribution of CNVs for the genetics of cardiomyopathy is limited. Reports in HCM include a Multiplex Ligation-Dependent Probe Amplification (MLPA) based study that failed to detect any CNVs in *MYBCP3* or *TNNT2* in a cohort of around 100 unrelated HCM patients [Bagnall et al., 2010]. Additionally, work on a single family identified a large *MYH7* deletion as the probable cause [Marian, 2012], which was detected using a PCR-based method and more recently another MPLA study found a single *MYBPC3* deletion [Chanavat et al., 2012] in a cohort of 100 unrelated genotype-negative patients.

Despite the fact that CNVs were only detected in a small percentage of our cohort, it raises the possibility that a patient with no identifiably causative variants can in fact harbour a structural variation not detected by direct sequencing. Consistent with this view, 3 out of 5 patients with confirmed CNVs did not carry any potentially causal variants in a sarcomeric or related gene, and a fourth patient only had a variant of unknown significance in *TNNT2*.

The patient with the most plausible single-nucleotide variant candidate (R495G in *MYBPC3*) also carries a CNV in the gene *LMNA*. This is an interesting finding because *LMNA* has traditionally been associated with DCM and not HCM [Vaikhanskaya et al., 2014; Pérez-Serra et al., 2015]. Furthermore, mutations in *LMNA* also cause Familial Partial Lipodystrophy 2 (FPL2), one of a group of heterogenous disorders that cause abnormal fat distribution. While DCM causing variants can occur across the gene, FPL2 variants are generally restricted to the the C-terminal [Lalitha Subramanyam, 2010]. Intriguingly, there have been reports of patients who present with FPL2 and are subsequently found to also have HCM [Araújo-Vilar et al., 2008; Chirico et al., 2014]. Information about the cholesterol levels or patterns of deposition and other FPL2 criteria were not available for the patients studied here. Analysis of such clinical data would be a natural way to investigate the potential role of *LMNA* in HCM further and to expand on recent work on refining phenotypes of a subset of these patients [Lopes et al., 2014].

Chapter 4

A novel method to deal with technical artefacts in exome sequencing data

4.1 Introduction

Population Stratification, Cryptic Relatedness (CR) and GC bias are three of many possible reasons why artefacts may exist in association studies based on High Throughput Sequencing data. This can confound the association between marker genotype and disease. Various methods have been implemented to account for this. One of the most commonly used involves performing Principal Component Analysis to identify orthogonal axes (PCs) which explain most genetic variation, which typically corresponds to PS and CR, then including the top PCAs as covariates in the regression to allow for any phenotypic variation caused by these sources. This can be incorporated into a PCA, thereby controlling for PS. Association studies that utilise pooled data, perhaps to increase study power, are more likely to suffer from technical artefacts/batch effects. These originate from the heterogenous nature of such studies, whether the samples can be grouped as cohorts that differ in their preparation, storage, sequencing technology etc. Spurious associations will arise when case/control ratios differ between cohorts; at the most extreme when some cohorts are all cases or all controls.

I introduce the concept of Technical Kinship (TK). TK is defined as a similarity matrix estimated on the SNP and INDEL missing/nonMissing matrix. Adapting the PCA approach of controlling for PS, we attempt to control for this by removing ten "technical" Principal Components in the LMM. This novel idea fails, highlighting the fact that technical bias in the data renders a more subtle effect than PS. By then using a LMM with a random effect with a correlation structure specified by the TK, I improve the ability to control for SNPs/INDELS that are more likely artefacts than true positives. This reduction of false positive inflation readily leads to more accurate association studies, thereby increasing the ability to identify disease causing genes. This analysis is performed on the UCL Exome Consortium of $\sim$ 4500 disease exomes which includes both Hypertrophic Cardiomyopathy (HCM) and Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) samples.

Different HTS machines and chemistries have different RD profiles but in general low coverage is a sign of low quality. This novel LMM is further refined by including a RD similarity matrix. The combination of this RD matrix and the TK matrix yields an improved result when compared to the standard linear regression with no such correction for artefacts.

4.1.1 Retinal Dystrophy - a motivating example

Retinitis Pigmentosa refers to a group of inherited retinal diseases that are characterised by photoreceptor and retinal pigment epithelium degeneration [Testa et al., 2014]. Symptoms typically include night blindness, visual field constriction and reduced electroretinographic waves (Figure 4.1). The gene Retinitis Pigmentosa GTPase Regulator (*RPGR*) is known to be responsible for $\sim$ 8.5% cases of the autosomal dominant form [Meindl, 1996; Churchill et al., 2013]. In a recent study we found a novel association between the gene Tubulin Tyrosine Ligase-Like family member 5 (*TTLL5*) and 28 individuals with "cone-first" retinal disease and clinical features that were atypical for ATP-binding cassette, sub-family A (*ABC1*), member 4 *ABCA4*-retinopathy [Sergouniotis et al., 2014]. *TTLL5* came second only to *RPGR* (Table 4.1). Two *RPGR* LOF variants, c.1586_1589delAGAG and c.401delT, a nonsense c.1627G>T and a missense c.1627G>A were found in the 88 cases examined (Figure 4.2).

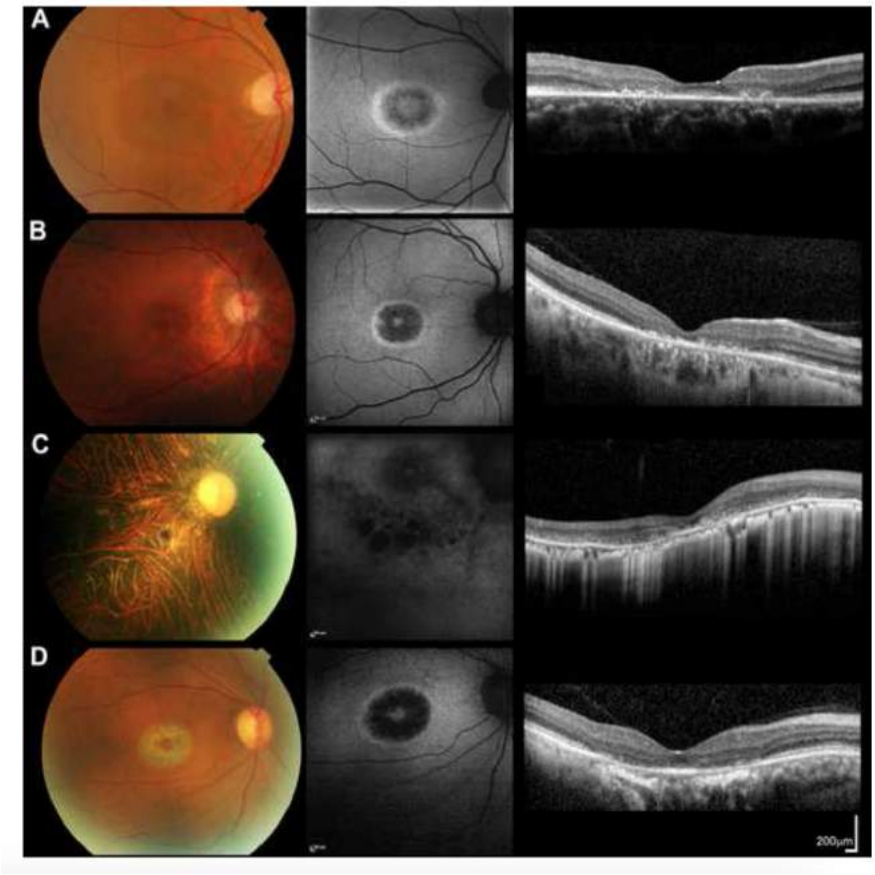


Figure 4.1: Color Fundus Photographs, Fundus Autofluorescence Images, and Foveal Optical Coherence Tomographs of the Right Eyes of Subjects CD1, CD2, CD3, and CD5. Images from subjects CD1 (aged 35 years; A), CD2 (aged 45 years; B), and CD5 (aged 53 years; D) are highly similar. Fundus autofluorescence imaging revealed a high-density concentric perifoveal ring surrounding irregular foveal autofluorescence in subjects CD1, CD2, and CD5; outside this ring, normal signal was observed (A, B, and D). In subject CD3 (aged 46 years; C), hypoautofluorescent patches were noted in the fovea and parafovea; this was combined with irregular autofluorescence outside the foveal region, suggesting more generalized retinal pigment epithelial dysfunction (C). Optical coherence tomography revealed abnormalities consistent with photoreceptor loss; they were either confined to the foveal region (subjects CD1, CD2, and CD5) or observed throughout the scan (subject CD3). Scale bars represent 200 m. Adapted from Sergouniotis,Chakarova, Murphy et al, 2014. I did not make this figure, included for illustration.

The support for these two genes is therefore well founded. However, when one looks at the third gene in the list, *C1R*, it becomes more difficult to verify its association. *C1R* is one of the proteases involved in the complement pathway, a vital part of the immune system [Rossi et al., 2014]. This fact alone fails to lend credence to it being thought of as a real association. Upon further examination of this gene, it was found that its entire signal was driven by the presence of the same LOF variant, chr12:7244369C>T, in 3 cases. As explained in detail in the Methods section (4.4.5), a PCA was performed on the missing/nonMissing

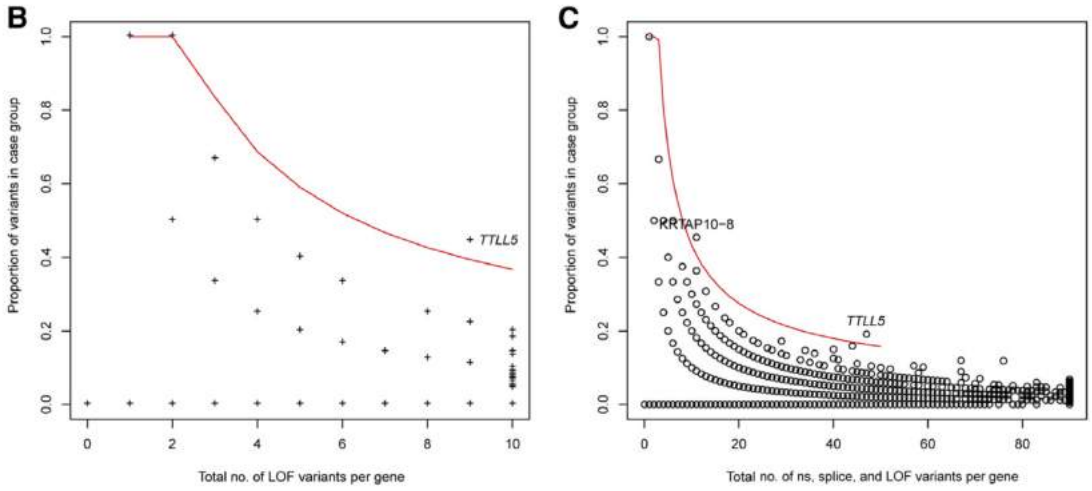


Figure 4.2: The number of presumed LOF alleles in both cases and controls (x-axis) and their proportions in the 23 retinal dystrophy samples (y-axis). The area above the red line is the gene based p-value threshold of $p \leq 10^{-}4$. I did not make this figure, included for illustration.

genotypes from UCL-ex to examine variant artefactual status. By including each of the PCs individually as covariates in separate linear regressions of phenotype on genotype, it was possible to determine that this variant is in fact strongly associated with the the 66th Principal Component (pvalue of 5.54e-25). The carriers of this SNP are outliers on the relevant PCA plot (Figure 4.3). It should be noted however that as Figure 4.4 illustrates, typically PCs after the 5th explain little of the total variance so the 66th is unlikely to influence the results significantly.

The *C1R* signal was thus declared an artefact. This filtering methodology is not the norm and differs largely from that employed by GATK (Section 1.8.3 on page 18). Be that as it may, it is an onerous approach as it required manually examining the most significant genes. A more thorough and efficient approach is therefore needed. That is one aim of this thesis.

Gene	Position	CaseCount	ControlCount	SKAT	Binomial
RPGR	chr23:38128893-38182760	4	2	0.000384659	2.57163e-06
TTLL5	chr14:76127372-76368547	4	5	0.000851088	2.05575e-05
C1R	chr12:7187848-7244382	3	3	0.000422697	0.000164897
OR5AU1	chr14:21623166-21624176	2	0	0.003364555	0.000420963
CDH3	chr16:68679283-68732274	2	0	0.003117345	0.000420963

Table 4.1: Top 5 Retinal Dystrophy candidate genes based on a binomial test for excess of variants in 23 cases compared to 1098 controls.

.

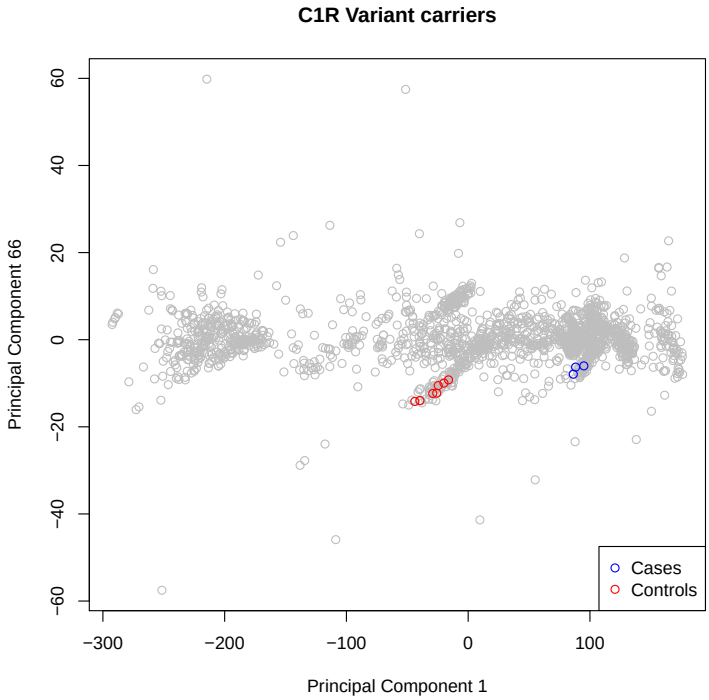


Figure 4.3: The first two principal components from the technical PCA. Highlighted are the locations of samples that contain the minor allele for the variant causing the false positive (chr12:7244369).

4.1.2 Crohn's Disease

The inflammation in Crohn's Disease (CD) is a transmural form of IBD (occurs across the entire wall of an organ). It can affect the entire gastrointestinal (GI) tract, from the mouth to the anus (Marshall et al., 2010). The affected regions may be discontinuous throughout the GI tract, and may locally involve strictures, abscesses or fistulas. Other symptoms include diarrhoea or constipation, abdominal pain, passing blood and signs of clinical obstruction (Baumgart and Sandborn, 2007). Diagnosis of CD is made by endoscopy and histology (Benevento et al., 2010). In addition to the incidence of CD varying from country to country, it also fluctuates between ethnic groups. The prevalence is 2-4 fold higher in people of Ashkenazi Jewish origin compared to non-Jewish Europeans [Kenny et al., 2012]. 800 such patients are included in UCL-ex.

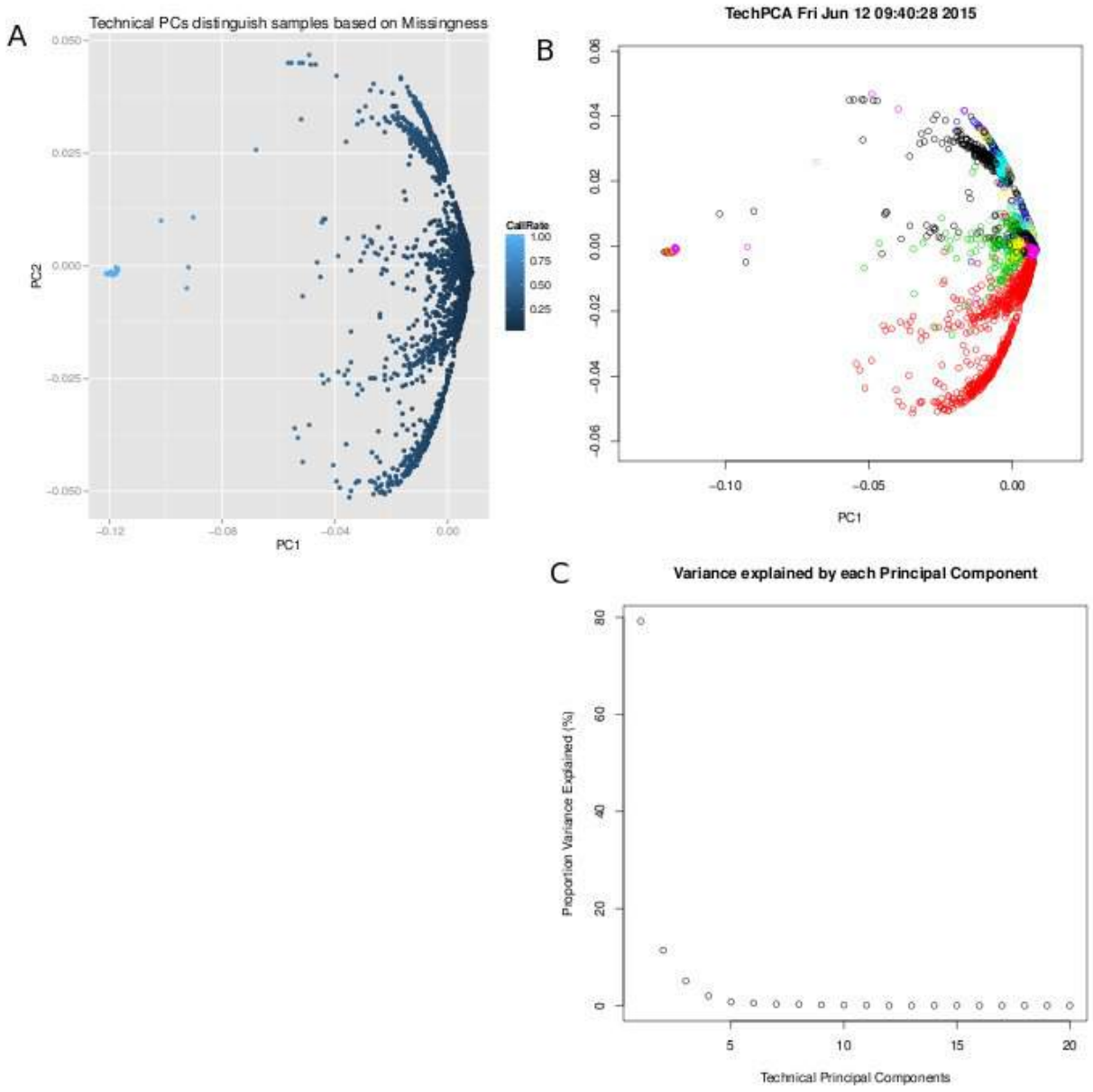


Figure 4.4: Principal Component Analysis of the Combined 1000 Genome Project and UCLex data for missingness estimation. (A) The samples are the dots, coloured on a scale from dark to light blue where the lighter the dot the higher the percentage of that samples SNPs that were not successfully called. (B) Same samples, but now they are coloured based on what research group they came from. (C) Scree plot showing the level of variance explained by each of the top 20 PCs.

4.1.3 Chapter aims

Dr. Vincent Plagnol created a pipeline that performs the initial alignment and variant calling of the UCL-ex. The work here builds on this, extending the pipeline to perform quality control, filtering and case control tests for all phenotypes within. This was all my work.

As mentioned already, HTS data can suffer from artefacts derived from many sources. These can be more apparent in a dataset such as UCL-ex where samples from multiple sources are pooled and all have rare diseases. This chapter is devoted to an attempt to create a novel statistical model that adapts classical methods from population genetics to try solve this problem. Dr. Plagnol and Dr. Doug Speed assisted with technical and statistical advice while the implementation and testing is my work.

4.2 Methods

4.2.1 UCL-ex Samples

Table 4.2 lists the breakdown of the samples, by number of samples and disease.

Phenotype	#Samples
Inflammatory Bowel Disease	799
Huntington's Disease	48
Ophthalmology	71
Ophthalmology	38
Ophthalmology	101
Ophthalmology	90
Ophthalmology	23
Ophthalmology	24
Ophthalmology	23
Dermatology	63
Sudden Cardiac Death	98
Keratoconus	12
Primary Immunodeficiency	128
Prion Disease	1112
Epilsepsy	164
ARVC	28
Bone Marrow Failure	184
Cone Rod Dystrophy	40

Table 4.2: UCLex Sample Information. Phenotype and number of samples

4.2.2 Data quality assessment

As might be expected, combining samples that have been prepared differently is not without some difficulty. For the most part, there is a high concordance rate. However, despite this, some SNPs will not be called in one or more methods (Figure 4.5A). The mean failure rate within each group was also examined (Figure 4.5B).

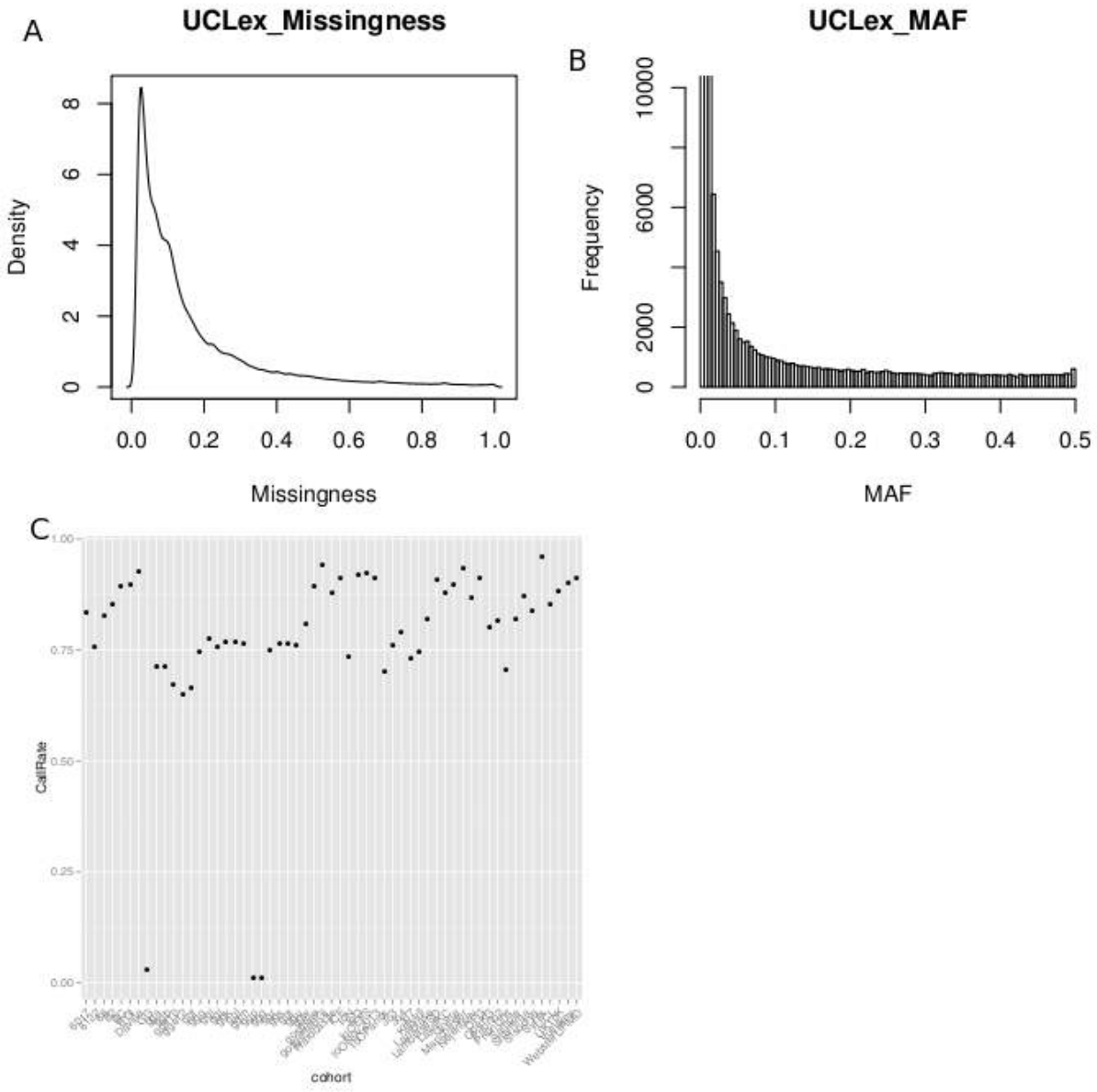


Figure 4.5: Genotyping call failure rate (A) across all samples within UCL-ex and (B) by group.

4.2.3 Attempting to identify samples with similar missingness patterns

Principal Component Analysis

A set of 5000 SNPs were chosen that are known to be well covered across all commonly used sequencing technologies. This was used because it would enable the PS control to be free from any bias associated with sample preparation differences. The 1000 Genome Project (1KG) provides a valuable resource of 1092 samples of known ancestry that have been sequenced with low-coverage genome and exome sequencing [Abecasis et al., 2012]. These data were combined with the UCL-ex samples at these 5000 loci. A PCA was then run on this subset. Generally, using the first two PCs is regarded as enough to adequately control for large scale PS [Price et al., 2006]. By comparing these first two PCs, one can readily see the separation of different populations (Figure 4.6). This PCA will be herein referred to as PC_{pop}.

In an attempt to identify any patterns of missingness in the data, the genotype matrix was converted to a missing/nonMissing matrix. Unlike PC_{pop}, this was performed on all SNPs as the sample preparation differences are of interest here. Regardless of exact genotype, if a SNP in a particular sample is called it was recoded as 1. If it was not called it was coded as 0. A PCA is then performed on this matrix to identify patterns of missingness (PC_{tech}). This can be visualised in the same way as the PC_{pop} (Figure 4.4A). As this shows, the first two PCs of this technical PCA can readily discriminate between samples that exhibit different missingness patterns. To understand this better, one can alternatively colour the samples based upon what research group they come from (Figure 4.4B). There is clear structure visible in the data. Removing samples based on this to create a more homogenous data set was attempted but the result is the removal of too many samples to be acceptable.

Adapting ADMIXTURE

ADMIXTURE is a model based clustering method that is used in population genetics to probabilistically assign samples to one of M populations, whether or not M is known. Traditionally, this is implemented on a matrix of sample genotypes and samples with similar haplotypes clustering together. As per the Pima Indian

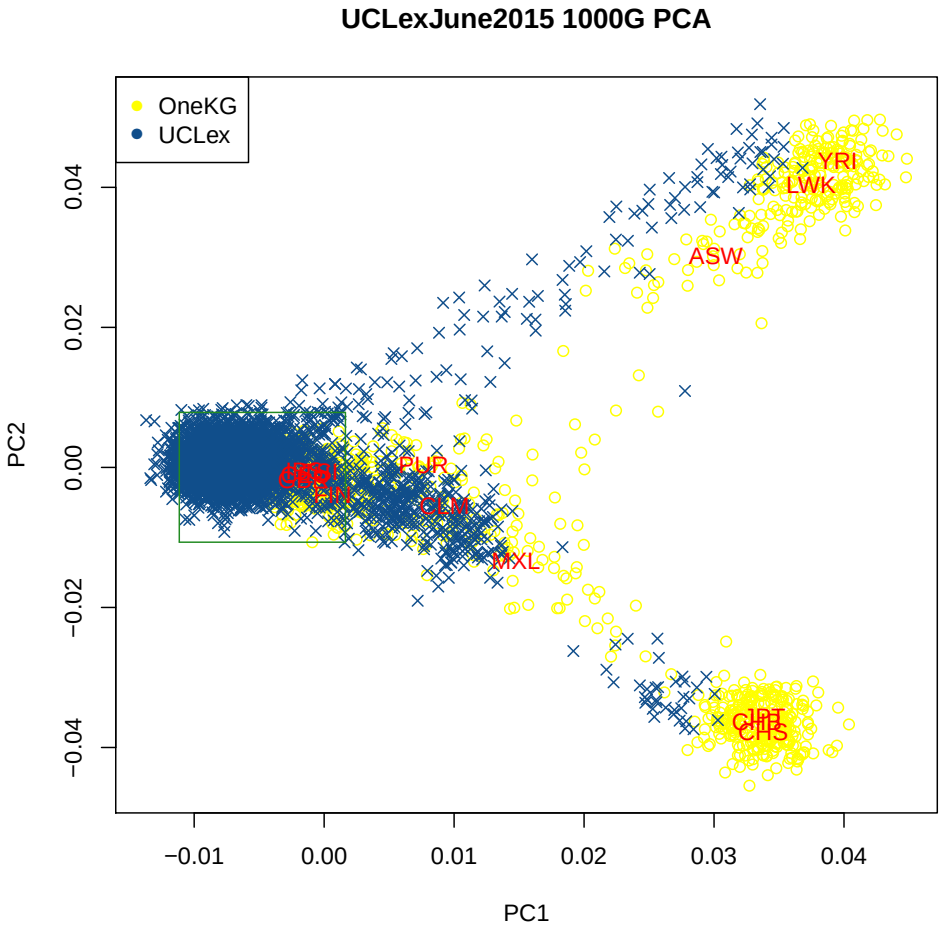


Figure 4.6: Principal Component Analysis of the Combined 1000 Genome Project and UCLex data for population estimation. The 1000G samples are yellow circles and the UCL-ex samples are shown as blue crosses. The coordinates for the different populations of the 1000G samples are shown. The green box demarcates the location of the Caucasian samples (CEU,TSI,GBR,IBS,FIN).

example mentioned in the introduction, restricting analyses to a closely matched population, as opposed to performing it on all samples can yield better quality data. The hypothesis here was that if it was possible to identify a group of samples (both cases & controls) that have a similar patterns of missingness, then this would allow for the calculation of more accurate case control association statistics.

To achieve this, the missing/nonMissing matrix was first converted to PLINK format [Purcell et al., 2007]. This was then supplied to ADMIXTURE, which was then ran 24 times, each time M was specified as a unique integer between 1 and 24. 24 was chosen as the maximum number of theoretical groups based

on the known number of batches of samples in UCL-ex (23). Figure 4.7 shows the estimated population assignation for M = 1:8.

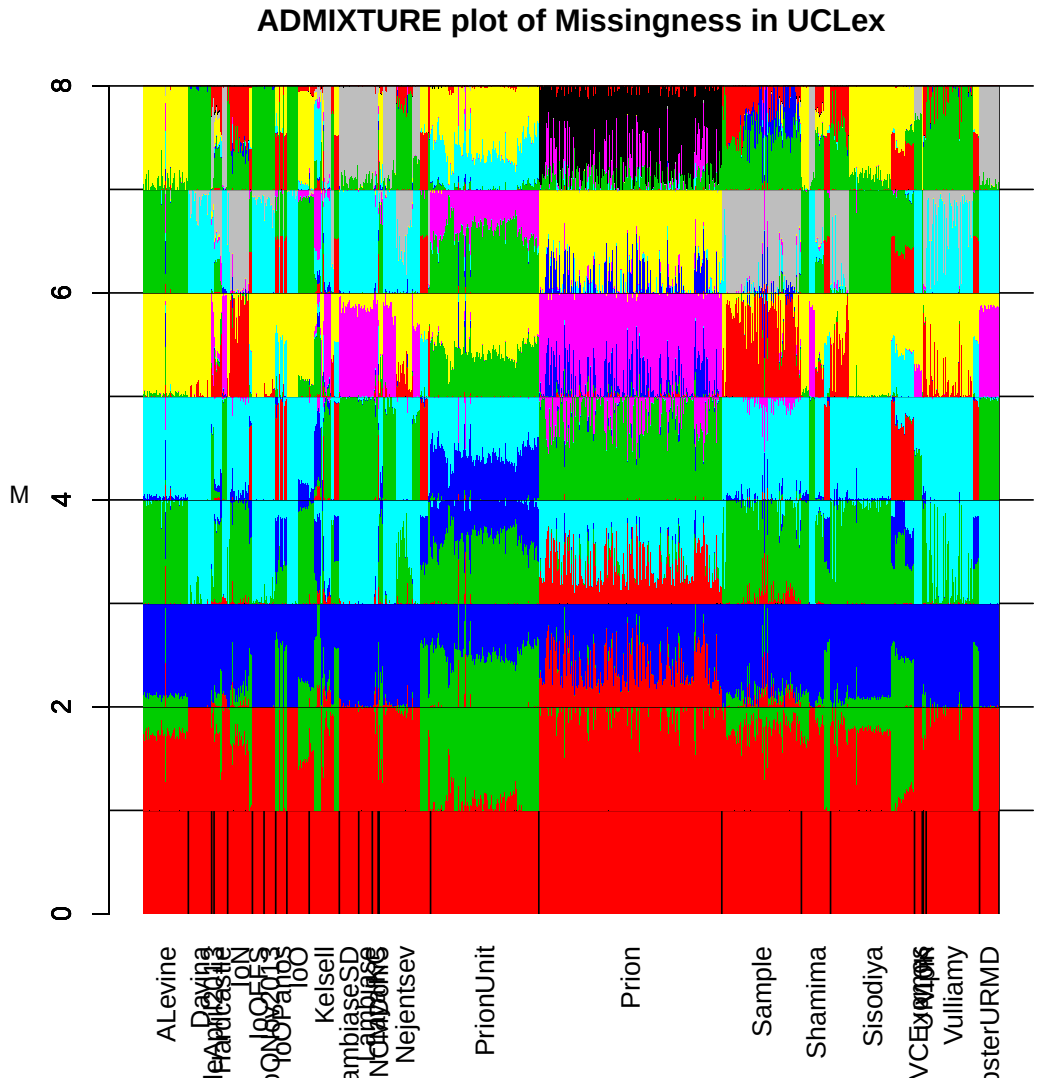


Figure 4.7: ADMIXTURE plot of the UCLex data illustrating the clustering of samples based on their missingness patterns. The y-axis shows the clustering of samples based on differing values of M, the sub-population limit for the ADMIXTURE algorithm. In each horizontal section, samples that are coloured the same are predicted by ADMIXTURE to have similar patterns of missingness. The individual samples are represented as vertical lines along the x-axis, with the grouping of samples based on their respective groups of origin labelled.

4.2.4 Mixed Model Association Testing

Imagine a given cohort comprised of distinct case and control samples. It is routine to imagine that sample preparations can differ between groups, at a rate higher than the within group variability (Figure 4.4). This can introduce confounding when one progresses to case-control association studies. This can present as not overly dissimilar to population stratification and cryptic relatedness, which are essentially the same confounder [Astle and Balding, 2009].

When testing SNPs for association with a phenotype, the basic linear model is most commonly used (Equation 4.1), where Y contains the phenotype, Z covariates, α fixed effects, X_j is the SNP being tested and β_j its effect size. The noise e is assumed to be normally distribution, $e \sim N(0, \sigma_e^2)$). This is typically solved using a score test, which estimates $\hat{\beta}_j$ and its standard error, then tests whether $\hat{\beta}_j$ is significantly non-zero.

$$Y = Z\alpha + \beta_j X_j + e \tag{4.1}$$

Generally, the covariates might include clinical factors such as age and sex, as well as often including top axes from PCA, in order to guard against population structure, as described above. In recent years, mixed model association testing has become more popular, where a random effect term is added to the basic linear model (Equation 4.2):

$$Y = Z\alpha + \beta_j X_j + g + e \tag{4.2}$$

g is a random effect, with distribution $N(0, K\sigma_g^2)$, where K is a specified kinship matrix, which is a measure of pairwise similarity across individual. Most commonly, $K = XX^T/N$, where the matrix X contains the standardized genotypes for the N SNPs, in which case we have Equation 4.3, where I is is the Identity matrix.

$$Var(Y) = \sigma_g^2 XX^T/N + \sigma_e^2 I \tag{4.3}$$

$$Y = Z\alpha + \beta_j X_j + \sum_{l=1}^{N} \gamma_l X_l e \quad with \quad \gamma_j N(0, \sigma_g^2/N) \tag{4.4}$$

Written this way (Equation 4.4), it becomes clear that mixed model analysis is equivalent assuming each SNP used when constructing the kinship matrix contributes to the phenotype with effect size γ_l. The random effect g is designed to pick up patterns due to PS and CR, and can also increase power by accounting for the contribution of causal variants away from the SNP being tested. Moreover, this approach avoids the need to decide how many PCs to include.

The aim of this chapter is to consider alternatives kinship matrices. Therefore, instead of representing genome-wide correlations across SNPs, we consider constructing K to reflect patterns of missingness and variance in RD.

4.2.5 Controlling for Read Depth

Another manifestation of the differing results from different capture technologies, or indeed from something as specific as the discrepancies between one lab's standard protocol to anothers might be a regional fluctuation in RD. For variant calling, particularly CNV identification, regional RD can be an important determinant of whether or not a call is made. However, most HTS technologies utilise a Polymerase Chain Reaction (PCR) amplification step, which introduces a bias in the library [Aird et al., 2010]. This skewed representation of reads can hinder accurate calling. A mostly effective way to control for this is to simply remove variants that have a depth below a given threshold. This is standard practise for many association studies [DePristo et al., 2011], with Picard (http://broadinstitute.github.io/picard/faq.html) being a widely used implementation. It has been recently shown though that this practise can introduce a bias, one that increases with RD [Zhou et al., 2014]. One goal of this project is therefore to attempt to refine Equation 4.3 to incorporate a correction

for RD. The logic is that one can create a 'Read Depth Kinship' matrix so that including it in the LMM will control for RD without the need for filtering. This is similar to the methods of calculating traditional kinships based on genotype in order to estimate and therefore control for ancestry.

4.2.6 Single Variants

To clarify, the models listed here are not all part of the final model used in the results section; they are instead included to discuss the process of model development.

These models were created and implemented with the help of Dr. Vincent Plagnol and Dr. Doug Speed.

Model 1 - Establishing a baseline with a standard Fixed Effect Linear Regression

Equation 4.5 was the basic model run for SNP j. Here, the covariates (Z) are PC_{pop}, the fixed effects Principal Components. Figure 4.8A illustrates the distribution of resultant pvalues. This clearly displays a false positive inflation, highlighting the need for correction.

$$Y = PC_{pop}\alpha_1 + \beta_j X_j + e \tag{4.5}$$

Model 2 - Adding Technical Principal Components to the Linear Regression

In the field of population genetics, the first two principal components are often used in a model to control for population stratification. The technical PCs were used here in a similar way to try to control for technical artefacts/bias. To that end, the top 10 PC_{tech}s were added as covariates into the model (Equation 4.6). This did not have a noticeable correction effect as shown by the pvalue distribution in Figure 4.9.

$$Y = PC_{pop}\alpha_1 + PC_{tk}\alpha_2 + \beta_j X_j + e \tag{4.6}$$

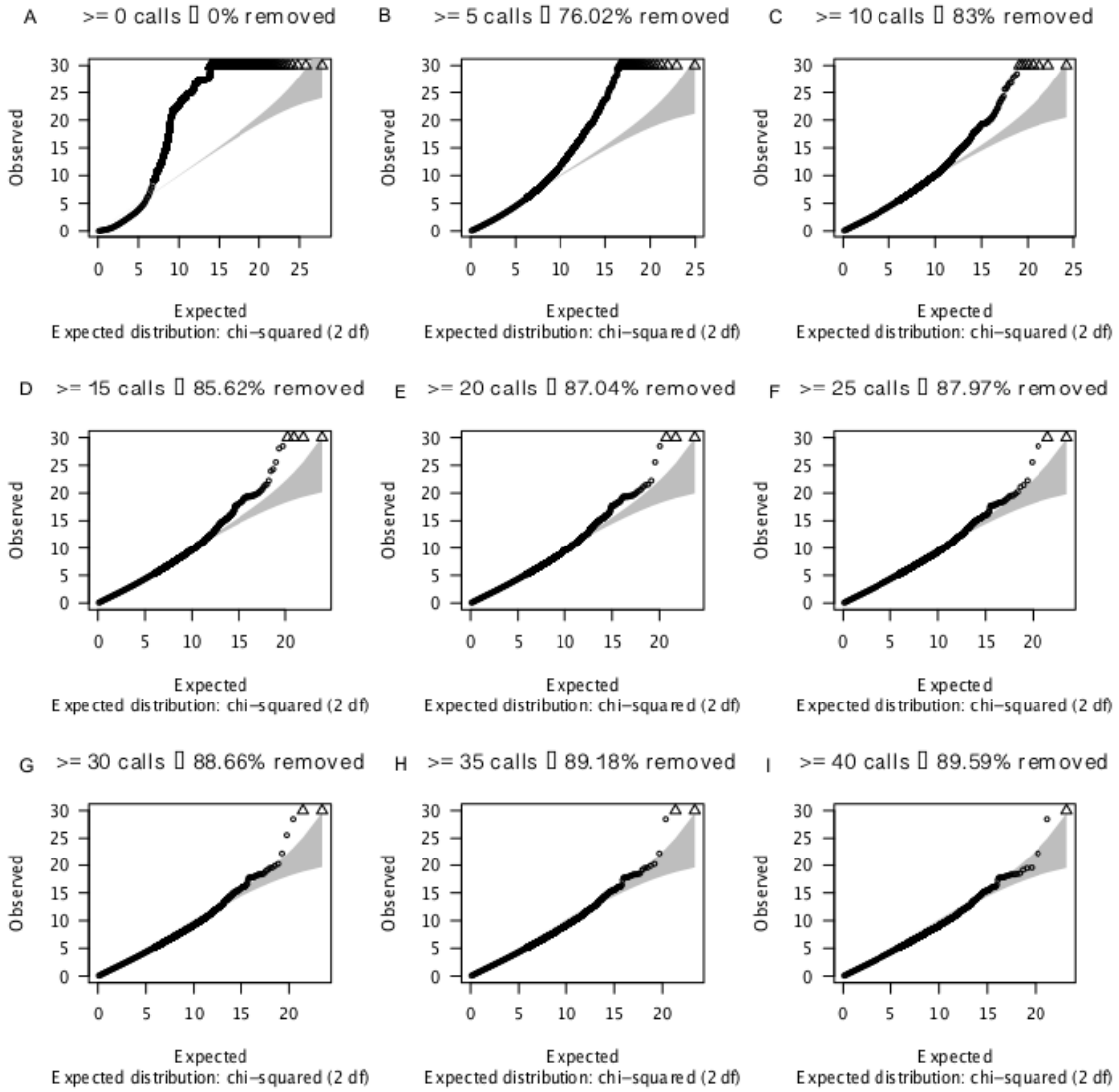


Figure 4.8: QQplots of Single Variant LMM with technical kinship correction on the PID cohort with the rest of UCL-ex as controls. The threshold of "common", in terms of observed counts of a particular variants minor allele, was varied to determine the least stringent useful cutoff.

Model 3- Linear mixed model with traditional kinship matrix.

As introduced in Section 1.7, SNP kinship matrices in LMMs can control for many confounders. Given that the PCs in the previous section did not work, I then tried Equation 4.7 which includes such a kinship matrix, g_{SNP}.

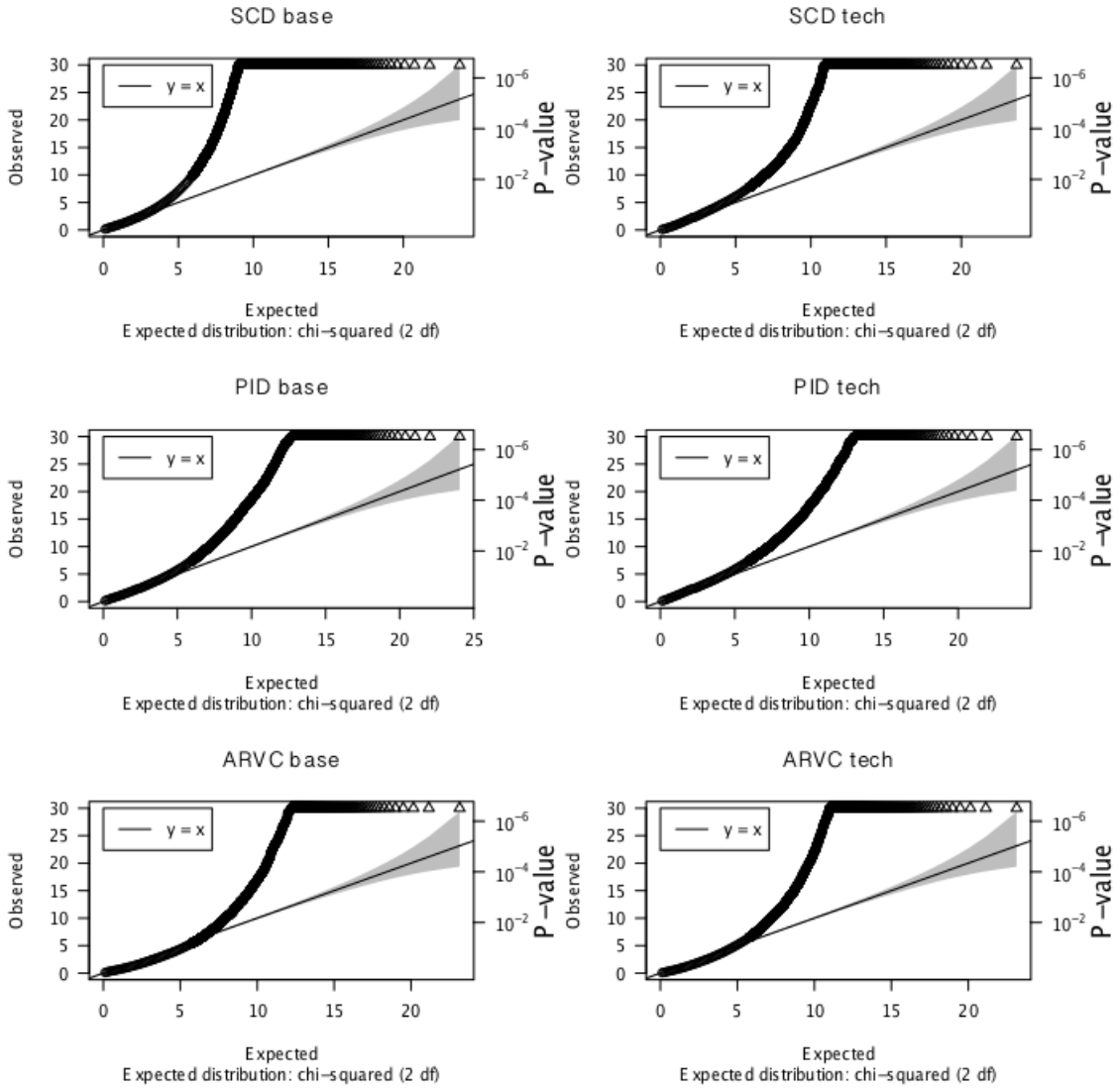


Figure 4.9: QQplots of Single Variant Linear Regression with ten technical Principal Components included to control for technical artefacts. This analysis was performed on the Sudden Cardiac Death (SCD) , Primary Immunodeficiency (PID) and Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) cohorts. The "base" QQplots include no artefact correction. These are compared to the "tech" models that do include these covariates.

$$Y = PC_{pop}\alpha_1 + \beta_j X_j + g_{SNP} + e \quad with \quad g_{SNP} \quad \sim \quad N(0, XX^T/N\sigma^2_{SNP}) \tag{4.7}$$

Model 4 - Adding the technical kinship matrix into the Linear Mixed Model

Essentially, Equation 4.8 differs from Equation 4.5 solely by replacing the traditional SNP kinship Matrix with the TK kinship previously described. The theory behind this is that SNPs that are artefacts will be explained by TK and therefore will not retain statistical significance as they will be controlled for. In theory, this should perform better than Equation 4.6 at correcting for artefacts as the kinship matrix will explain all the variability attributable to missingness.

$$Y = PC_{pop}\alpha_1 + \beta_j X_j + g_{TK} + e \quad with \quad g_{TK} \quad \sim \quad N(0, TK\sigma^2_{TK}) \tag{4.8}$$

Model 5 - Addition of Read Depth Kinship Matrix.

The final model builds upon Equation 4.8 by adding a RD Kinship Matrix to further control for the data artefacts. The log of the raw RD values was used to gain a more sensible representation and in an attempt to reduce its correlation with TK. Ten SNP PCs and five Hapmap PCs were further included to eliminate PS, creating the final model shown as Equation 4.9. SNPs that had a MAF of $\geq 1\%$, missingess rate of $\leq 20\%$ and a Hardy-Weinberg Equilibrium (HWE) pvalue of ≥ 0.001 were kept for this analysis.

$$Y = PC_{pop}\alpha_1 + \beta_j X_j + g_{TK} + g_{RD} + e \quad with \quad g_{TK} \quad \sim \quad N(0, TK\sigma^2_{TK}) \quad and \quad g_{RD} \quad \sim \quad N(0, RD\sigma^2_{RD}) \tag{4.9}$$

4.2.7 Computational cost considerations

As discussed in Section 1.7, exact solving of the mixed model for each SNP is computationally feasible with just one kinship random effect. However, we progress to situations with more than one kinship such as Equation 4.9 in which case it is necessary to use the approximation used by GRAMMAR.

Performing a GWAS on a dataset the size of UCL-ex, with 4500 WES' at the time of writing, is

computationally demanding. The burden of such LMMs are well known in the literature, as the computation time increases to the scale of n^3 Zhang et al. [2010]; Yu et al. [2006]. This was mentioned in Section 1.7. Throughout this thesis, PCs were used instead of full Kinship matrices where possible as they are far less computationally demanding while still offering adequate correction for PS.

LMMs do however offer the ability to correct for population stratification and cryptic relatedness alongside NGS artefacts by Variance Component Estimation (VCE). VCE has a long history in genetics, from its origins in animal breeding to Quantitative Trait Loci (QTL) analysis [Amos, 1994; Almasy and Blangero, 1998]. Estimating these parameters is intensive in its own right, as iterations are required for each marker [Gilmour, A; Thomson, R; Cullis, 1995]. 2007 saw the introduction of GRAMMAR, an expedited solution to this problem [Aulchenko et al., 2007]. This combines a mixed model analysis with a basic linear regression. This divides the analysis into at least two steps; firstly the VCE without marker data. The residuals from this step are then used as a novel phenotype for a classical association test with linear regression. In our case, the initial step uses Restricted Maximum Likelihood (REML) to estimate the variance explained by two Kinship matrices, "RD" and "TechnicalKinship", and any population stratification parameters that are additionally included (Equation 4.10). The residuals from this, now free from artefacts associated with RD and informative missingness, are used in Equation 4.11.

$$Y = Z\alpha + \quad g_{RD} + \quad g_{TK} + \quad e \tag{4.10}$$

i.e. Equation 4.4 without the SNP X_j, then use the resulting estimates of the fixed and random effects, $\hat{\alpha}$ and $\hat{g}$ respectively, to compute the "residual phenotype" (Equation 4.11). This phenotype is compared against the original phenotype in Figure 4.10. Finally, the residuals Y^* can be used as the phenotype in Equation 4.1.

$$Y* = Y - Z\hat{\alpha} - g_{\hat{RD}} - g_{\hat{TK}} \tag{4.11}$$

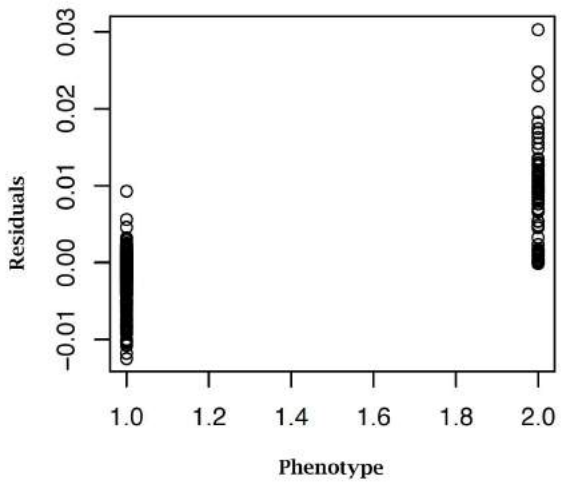


Figure 4.10: Comparison of phenotype to its residuals for a given trait in UCL-ex. The X axis represents the case control (1/2) phenotype while the Y axis is the 'Residual Phenotype' as discussed for use in Equation 4.11.

4.2.8 Gene based tests

Variants with a MAF of $\leq 1\%$ are not tractable to single variant approximations such as Equation 4.6 due to a lack of statistical power to detect a signal. Various methods were used during this work in an attempt to glean sensible data from these less tractable variants. Different forms of Gene based tests were used. Here, the hypothesis was that one could first remove variants that are associated with the technical PCs, $PC_{tech1:10}$. This would remove variants that are thought to be technical artefacts. To do so, each variant was regressed as in Equation 4.12. This is an exclusion test, as a significant pvalue (p <0.0001) indicates that there is a significant association between the variant and $PC_{tech1:10}$. Therefore such associated variants were removed from further analysis.

$$SNP \sim Y + PC_{pop1:2} + e \tag{4.12}$$

$$SNP \sim Y + PC_{pop1:2} + PC_{tech1:10} + e$$

Rare (MAF <0.3%) and non-synonymous, LOF or splicing variants were retained from the variants

that remained after the PC_{tech} filtering of Equation 4.12. MAF was defined separately based on the 1000G samples and on a random quarter of the UCL-ex controls (which were not used for subsequent analyses). This filtered list of variants was then subjected to both SKAT and a basic binomial test that tests for an excess of variants in cases compared to controls.

4.3 Sudden Cardiac Death

4.3.1 SCD-UCLex Single Variant Association Tests

68 samples diagnosed with Sudden Cardiac Death were included in the UCL-ex consortium. When one has a low number of cases like this, an improperly designed study may remove any possibility of retaining enough power to detect SNPs of weak or moderate effect. In general, case control studies using unrelated samples have more power than family based studies, in part due to the increased ease of obtaining large numbers of samples [Risch and Teng, 1998]. However it has been shown that including families with multiple affected siblings in a case control of mostly unrelated individuals can further increase the power because it enriches the study with disease alleles [Risch, 2000]. This may be more true in polygenic diseases such as SCD than in monogenic diseases [Li et al., 2006]. Much work has been done to try identify the optimal study design, using different combinations of classical tests such as the Transmission Disequilibrium Test [Spielman et al., 1993] with linkage and association studies [Fingerlin et al., 2002]. This was further improved with likelihood based strategies, such as a combined likelihood approach that multiplies the likelihood contributions of families and unrelated samples together [Nagelkerke et al., 2004]. Even more recently, it has been shown that by combining aggregated haplotype weighted counts from case control and trios under a generalised linear model, you can have a more powerful and cost effective study than other version alone [Wen and Tsai, 2014].

This cohort includes a family whose pedigree is shown in Figure 4.11). 2 case control tests were performed on this cohort, one with the family excluded (everyone except the proband) (Table 4.3) and one with them included (Table 4.4). This was done to ascertain if inclusion of family members substantially

increased the power to detect real signal and did not simply highlight the existence of private mutations of no clinical consequence. The related QQ plots are in Figure 4.12. This does not display a good correction for PS with a large level of inflation remaining so the results may be artefactual.

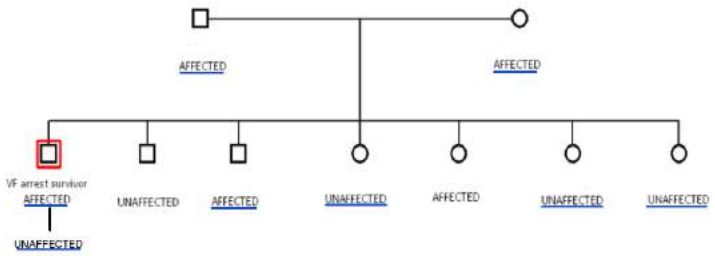


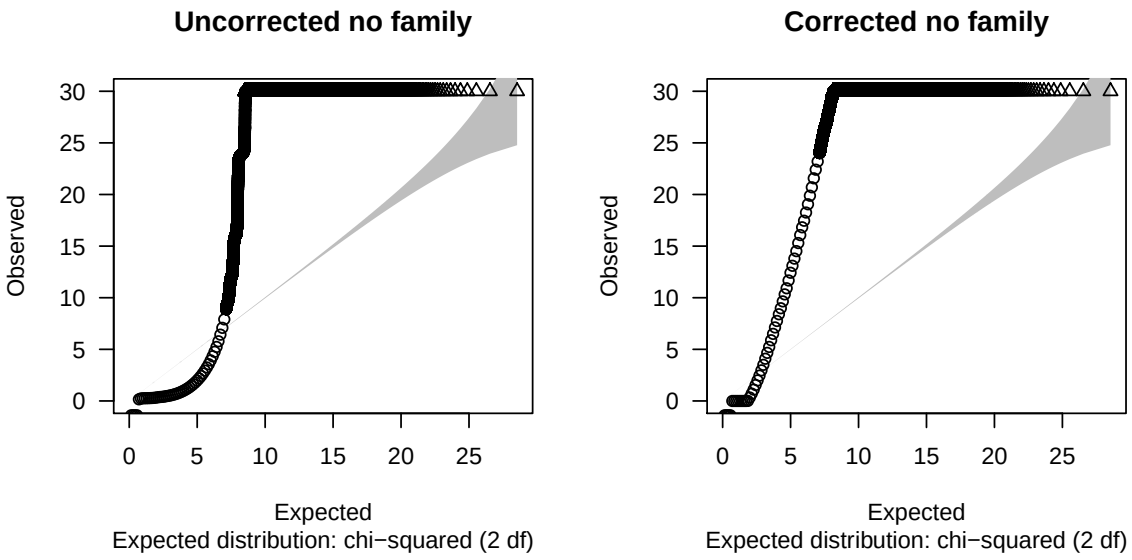
Figure 4.11: Pedigree of the J wave Family discussed in relation to Table 4.4. A blue line underneath the phenotype indicates that this sample was present in this study. The sample highlighted with the red box is the proband

rsID	SNP	Gene	Fisher	LRp	LMMp	OR	#Hom.SCD(#n62)	#Hom.ctrl(#n4268)	#Het.SCD	#Het.ctrl
NA	c.2084_2107del	C10orf71	1.74E-15	1.00E-16	1.00E-16	1.73E+01	6	22	7	8
NA	c.2093_2094insCACACG	C10orf71	1.05E-16	1.00E-16	1.00E-16	2.03E+01	6	22	7	8
NA	c.185C>T	OR10G4	2.24E-02	3.32E-01	1.51E-11	8.91E+01	0	4	1	34
rs11538191	c.-398C>T	C12orf44	1.00E+00	1.37E-02	4.56E-09	0.00E+00	0	3	0	18

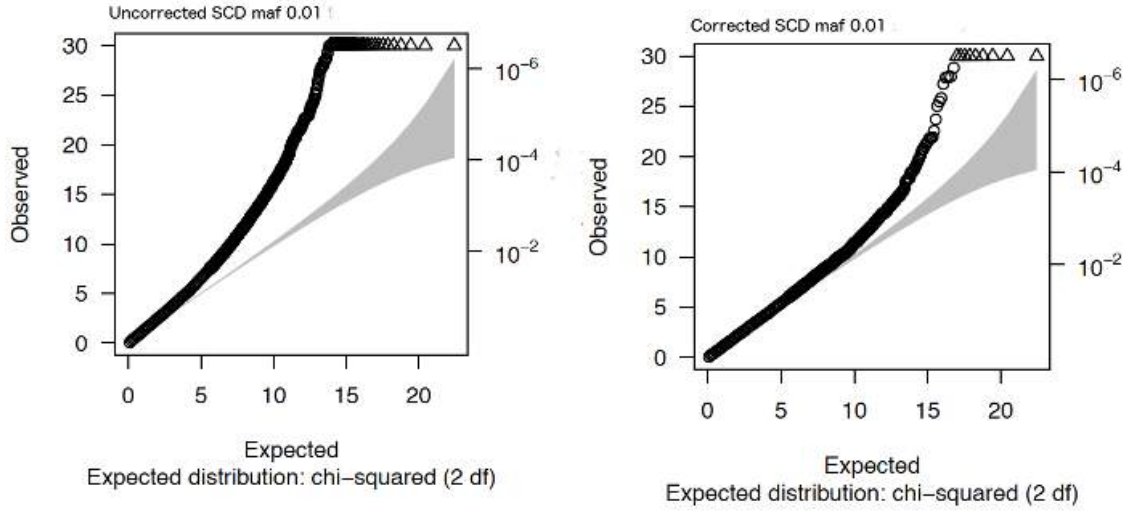
Table 4.3: Sudden Cardiac Death(SCD) Single Variant Results without Jwave family. SNP details the position of the tested variant (hg19). Gene is the HUGO name for the gene in which the SNP resides. FisherP is the pvalue from Fisher's exact test. LRp is the Linear Regression pvalue with no covariates or kinship matrices. LMMp is the pvalue from Equation 4.9. OR is the risk odds ratio. 'Homs' are homozygotes for the minor allele, while 'Hets' are heterozygotes

rsID	SNP	Gene	Fisher	LRp	LMMp	OR	#Hom.SCD(#n68)	#Hom.ctrl(#n4268)	#Het.SCD	#Het.ctrl
rs141832071	c.3463C>G	*FOCAD*	$9.17 * 10^{-4}$	$\leq 1. * 10^{-16}$	$\leq 1. * 10^{-16}$	69	0	0	2	3
NA	c.209C>T	*ZNF323*	$4.51 * 10^{-19}$	$\leq 1. * 10^{-16}$	$5.52 * 10^{-13}$	NA	1	0	7	0
NA	c.237G>T	*ZNF323*	$4.46 * 10^{-19}$	$\leq 1. * 10^{-16}$	$6.14 * 10^{-13}$	NA	1	0	7	0
NA	c.781C>T	*OR5V1*	$6.41 * 10^{-19}$	$\leq 1. * 10^{-16}$	$7.03 * 10^{-13}$	NA	1	0	7	0
rs11466802	c.2413G>A	*ADAM19*	$7.29 * 10^{-4}$	$\leq 1. * 10^{-16}$	$3.23 * 10^{-12}$	90	0	0	2	2
NA	c.591A>G	*FSTL1*	$6.17 * 10^{-15}$	$\leq 1. * 10^{-16}$	$4.01 * 10^{-12}$	NA	0	0	7	0
rs146280894	.285G>A	*CEP97*	$6.55 * 10^{-15}$	$\leq 1. * 10^{-16}$	$5.4 * 10^{-12}$	NA	0	0	7	0
rs376775426	c.333C>T	*BTG2*	$7.32 * 10^{-15}$	$\leq 1. * 10^{-16}$	$7.18 * 10^{-12}$	NA	0	0	7	0
rs267603590	c.306G>A	*HSD17B6*	$1.26 * 10^{-12}$	$\leq 1. * 10^{-16}$	$8.8 * 10^{-10}$	NA	0	0	6	0

Table 4.4: Sudden Cardiac Death(SCD) Single Variant Results with Jwave family. SNP details the position of the tested variant (hg19). Gene is the HUGO name for the gene in which the SNP resides. FisherP is the pvalue from Fisher's exact test. LRp is the Linear Regression pvalue with no covariates or kinship matrices. LMMp is the pvalue from Equation 4.9. OR is the risk odds ratio. 'Homs' are homozygotes for the minor allele, while 'Hets' are heterozygotes.



(a) These plots illustrate the case control analysis of only unrelated cases and controls



(b) The J wave family is included in the case control analysis shown here.

Figure 4.12: Sudden Cardiac Death (SCD) single variant mixed model association results. QQplots of the uncorrected (a) and corrected (b) SCD analysis for variants with a MAF of $\geq 1\%$ are shown.

4.3.2 An enhanced model for gene based correction of technical artefacts

As mentioned previously, Equation 4.9 on page 81 worked well for at least some traits and for variants with a MAF of $\geq 1\%$. This left rare variants uncorrected. Power to detect and correct rare variants would be gained by pooling variants into a region based testing procedure. For this, grouping variants based on what genes they lie in seems intuitive biologically. For all sequenced genes, the SCD samples were compared to the rest of UCL-ex with SKAT and a Binomial test (further methodological details in Section 4.2.8). Figure 4.13 shows that this region-centric approach improves the distribution over the single variant scores shown in Figure 4.12. The most significant genes, ranked by Binomial pvalue are shown in Table 4.5. This table is dubious as none of these genes have been reliably associated with SCD previously: The SCD cohort remains difficult to interpret, so a different approach is needed.

Gene	Position	Case Counts(n=90)	Control Counts(n=2,236)	SKAT	Binomial
OR5V1	chr6:29323076-29323905	10	7	3.011447e-15	1.831340e-12
PCDHGA9	chr5:140782689-140784943	13	34	1.934987e-10	3.476049e-09
PHKA1	chrX:71800901-71933724	11	1	1.430687e-10	1.212894e-08
ZNF280A	chr22:22868366-22869937	15	31	2.629286e-10	1.496724e-08
RSAD2	chr2:7017943-7027279	9	13	5.225212e-14	2.682962e-08

Table 4.5: Top 5 Sudden Cardiac Death candidate genes based on the binomial test. The criteria for retaining variants are: GATK Variant Quality score of PASS, MAF (MAF of $\leq 0.3\%$) , $\leq 10\%$ missingness across all samples, non-synonymous, LOF or affecting splicing.

Methods

In an effort to increase the improvement, two alternative methods, similar to each other, were implemented. The single variant permutations of Figure 4.14 reveal the power of this approach. To create a null distribution based on our dataset, phenotype status for all samples was permuted using the software LDAK (v3.0). This entails retaining the same number of cases and controls, but altering randomly which samples are assigned as cases and controls, respectively. While such permutations will not improve the low resolution caused by a small sample size, it will remove the technical artefacts that are associated with case control status. A 100 such permutations were run to create a null distribution relevant to the data. This was compared to the real pvalue from both the basic LMM and Equation 4.9. The QQplots from a random permutation, and both

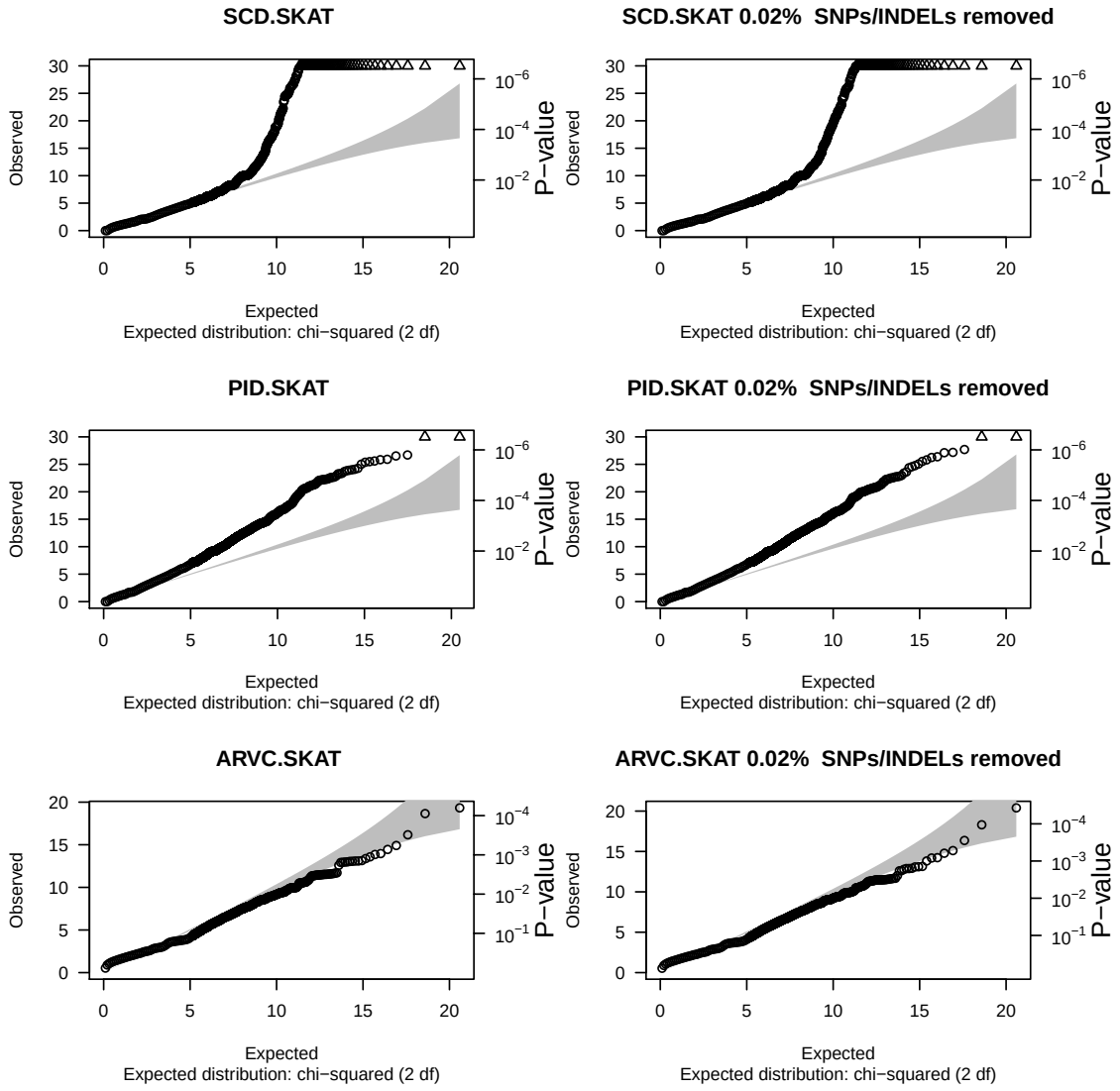


Figure 4.13: SKAT Gene based tests for the PID, ARVC and SCD cohorts. Each circle is the gene based pvalue from the SKAT test. QQplots compare the observed distribution of pvalues to the expected Chi Squared distribution. Before the pvalue for each gene is calculated, some variants are filtered/removed. On the left graphs, the criteria for retaining variants are: GATK Variant Quality score of PASS, MAF (MAF of $\leq 0.3\%$) , $\leq 10\%$ missingness across all samples, non-synonymous, LOF or affecting splicing. (B) For the graphs on the right, the same criteria are used but additionally variants are filtered based on the technical PCA. The first ten principal components (PC) are included in the linear regression as covariates. SNPs that are associated with the technical PC are removed. The percentage of SNPs/INDELs removed across all genes is included in the figure titles.

non-permuted tests are shown in Figure 4.15. As this shows, Equation 4.9 goes some way to correcting the test, resulting in data that is somewhat interpretable.

TK was included in the mixed effects model of Equation 4.4. Here, TK is calculated only on the

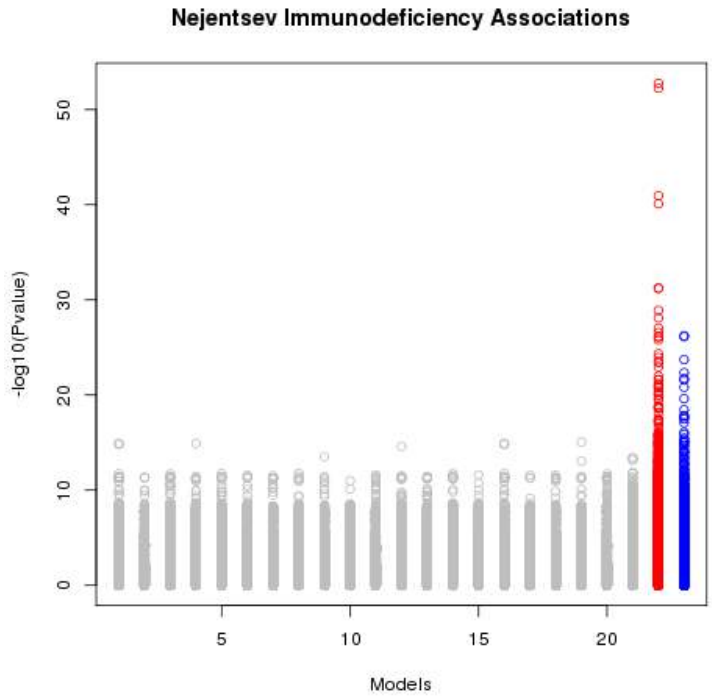


Figure 4.14: Association results for SNPs/INDELs with PID. The -log_{10} (y-axis) of the pvalues from 20 permutations (grey), Linear Regression (red) and Linear Mixed Model with technical kinship (blue).

SNPs in the gene of interest. The likelihood ratio statistic $-2log(L(Y|\hat{\sigma}_e^2, \hat{\sigma}_g^2, TK)/L(Y|\hat{\sigma}_e^2, 0, TK)$ has an approximate null distribution χ^2(Equation 4.3)/2. σ_e^2 is calculated individually for the numerator and the denominator. Restricted Maximum Likelihood (REML) then derives the model likelihood from Equation 4.3, an efficient process when you have more SNPs(N) than individuals (n). On a gene based level however, n is typically greater than N, so to expedite model likelihood calculation, Equation 4.4 is used to abrogate the need for K calculation (Speed *et al*, in preparation).

To identify candidate genes from this, some pvalue comparisons were made. Firstly, a gene is unlikely to be truly disease causing if it has a pvalue within the range of pvalues seen in the permutations. Table 4.6 therefore includes the minimum permuted pvalue for each gene. *LRRC37A2* has an uncorrected pvalue of 2.40e-30, but the corrected pvalue is less extreme than the permuted pvalue, thereby rendering it a false positive. The gene Phosphodiesterase Interacting Protein 4 (*PDE4DIP*) has an uncorrected pvalue of p <1e-40. Phosphodiesterases regulate cyclic nucleotide signalling, and are therefore of clinical importance

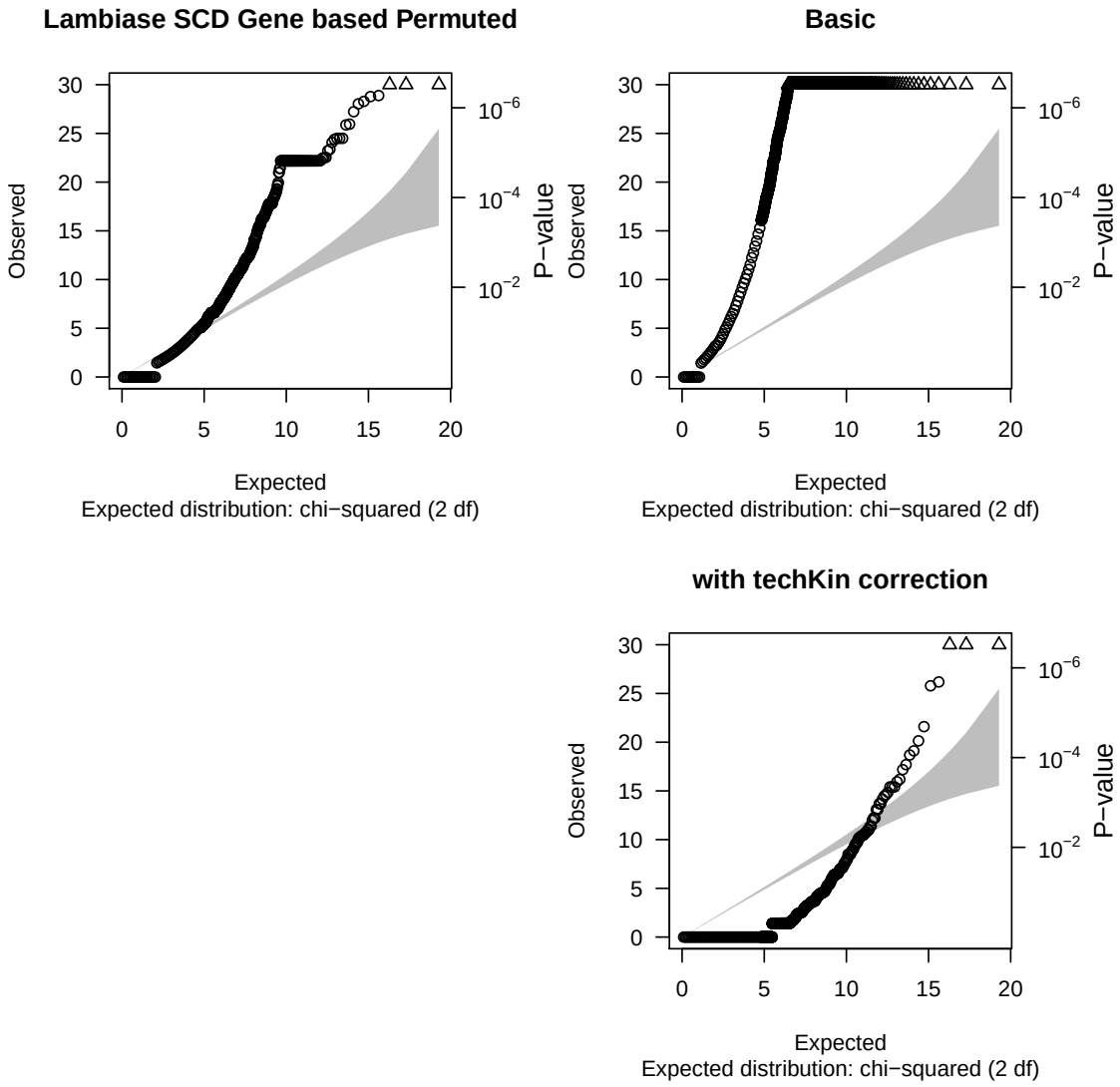


Figure 4.15: Comparing the distribution from different Sudden Cardiac Death Gene based tests.

[Jeon et al., 2005]. *PDE4DIP/MMGL4* has been reported to phosphorylate *MYBPC3* [Uys et al., 2011]. Variations in *MYBPC3* are known to confer an increased risk to developing Hypertrophic Cardiomyopathy. While this would place *PDE4DIP* as a likely novel gene for SCD risk, the pvalue when TK is included changes to 1 (while the minimum permuted pvalue is 7.97e-05). Even if *PDE4DIP* is in actuality a disease causing gene for SCD, this corrected pvalue of 1 means that any real signal correlates strongly with the batch effect removed by TK. To determine if it is a real signal or a false positive, the batch effect would have to be non existent. This could be achieved, for example, by preparing cases and controls in entirely the same

fashion. This would include everything from DNA extraction, to sample storage all the way to sequencing and processing. The same can be said for *RIMS3*. The remaining genes in this table retain significance with the inclusion of TK, meaning that it corrects for batch effects. When their permuted pvalues are less significant than the corrected pvalues, then that is evidence for a true association.

Gene	Position	Pvalue.no.correction	Pvalue.with.correction	Min.Permuted.pvalue
SPACA5B	chr23:47990038-47991995	1	1.26e-08	3.33e-06
FAM58A	chr23:152853382-152864632	0	1.26e-08	4.26e-05
SSX6	chr23:47967366-47980068	1	5.88e-08	1.18e-06
RIMS3	chr1:41086351-41131324	1	2.06e-06	1.11e-07
PROKR2	chr20:5282685-5295015	3.94e-01	2.51e-06	2.93e-03
LRRC37A2	chr17:44590075-44633014	2.40e-30	2.04e-05	1.91e-06
PDE4DIP	chr1:148889463-149033016	<1e-40	1	7.97e-05

Table 4.6: Top 5 Sudden Cardiac Death candidate genes based on the gene based technical kinship corrected pvalue. 98 cases were compared to 4,236 controls.

4.4 Results

4.4.1 Initial data quality assessment

The 4334 exomes were stored in a 4334 * 884887 matrix (Samples * variants). All variants were either exonic or altered splicing. An initial quality check examined the call rate across all SNPs (Figure 4.5). As expected, the vast majority had a failure rate of <20%. The call rate varied from group to group however, over a range of 2-25% (Figure 4.5B).

4.4.2 Principal Component Analysis

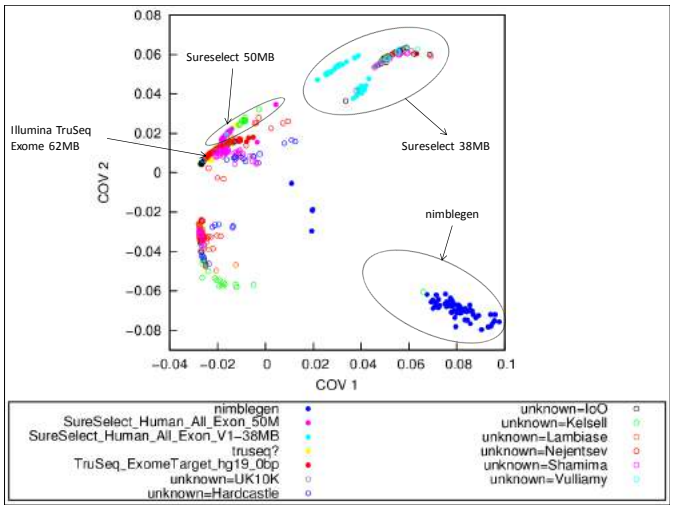
A PCA was performed on the $\sim$ 5000 SNPs that are known to be well covered across all commonly used sequencing technologies. This was used because it would enable the Population Stratification (PS) control to be free from any bias associated with sample preparation differences. The first two Eigenvectors of this PC_{pop} were readily able to discriminate population substructure in UCL-ex by comparing it against the samples of known ethnicity from the 1000G project (Figure 4.6). While the majority of UCL-ex was determined to be of Caucasian origin, as expected, some were more likely African or Asian. Such population substructure was

controlled for in the association testing used here by including the top PCs as covariates in both the Linear Regression and the Linear Mixed Model.

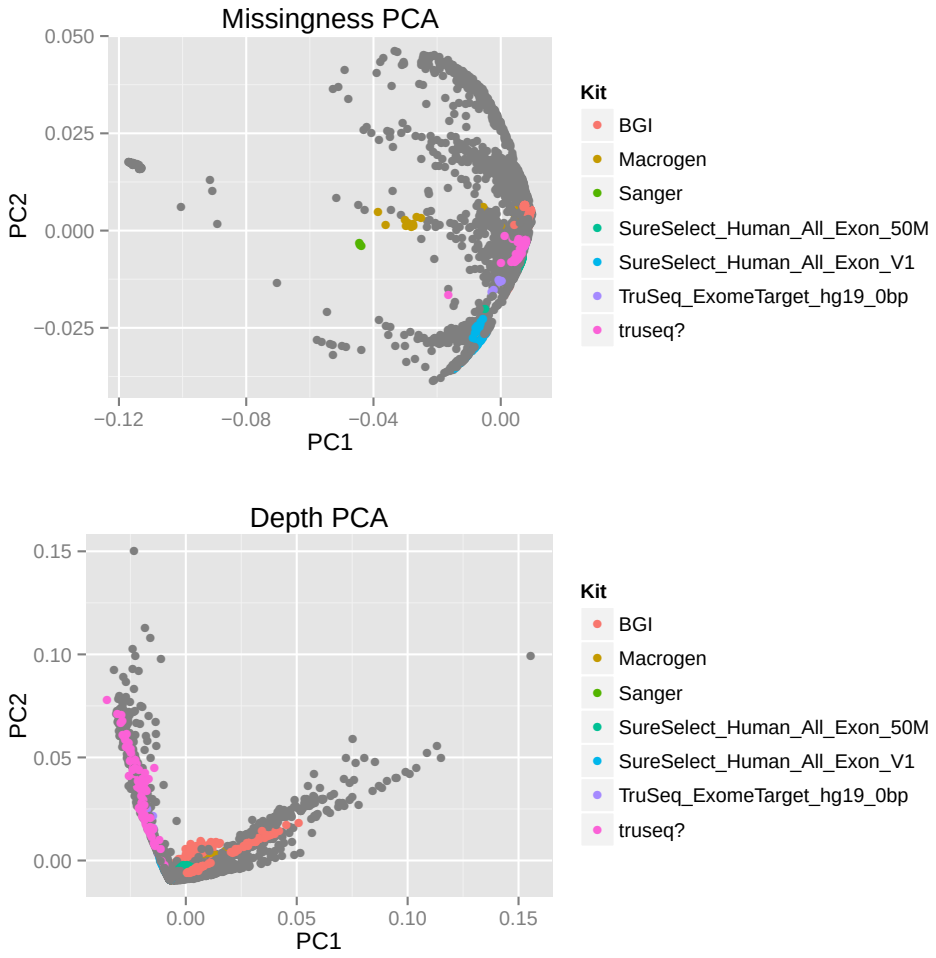
In addition to the PC_{pop} approach, a PCA was performed on the missing/nonMissing matrix of all variants (PC_{tech}). Figure 4.4A shows the first two PCs' ability to differentiate samples based on their patterns of missingness. The samples are the dots, coloured on a scale from light to dark blue where the darker the dot the higher the percentage of that samples SNPs that were not successfully called. The general trend from this is that samples with similar numbers of NA SNPs/INDELs cluster together. Figure 4.4B is the same plot except for the fact that samples are coloured based on their research group of origin. This reinforces the idea that technical artefacts can be highly associated with case control status. Figure 4.16A illustrates this further by including just the samples whose sequencing chemistries are known. Samples from different traits readily cluster together when viewed by Technical PCs. Figures 4.16B and C show how by overlaying these well characterised samples across all of UCL-ex, you can reliably predict the HTS platform used. The same method was applied to a RD matrix. Figure 4.17 shows the PCA plot from this. It is relatively uninformative as it does not offer much discrimination. The Scree plot in the lower section of this Figure shows that there is little variance ($\leq 5\%$) explained by PCs and below. By comparing these Figures, it shows that read depth is not as useful a determinant as missingness in identifying clusters in the data.

4.4.3 ADMIXTURE based sample separation

ADMIXTURE, the model based clustering approach to identifying population stratification, was used here to see if it was possible to identify a group of samples (both cases & controls) that have a similar patterns of missingness. This may then allow for the calculation of more accurate case control association statistics by filtering controls so that they matched the cases as closely as possible. This programme was run numerous times; each time the parameter M that governed the desired number of subpopulations for ADMIXTURE to resolve was varied from 1 to 25. Values of $M \geq 4$ start to show signs of empty resolution (Figure 4.7). Similar to Figure 4.4B, this ADMIXTURE plot highlights the Prion samples as outliers. A case control study that naively used all samples in UCL-ex could be affected by this grouping. Variants may be erroneously called as

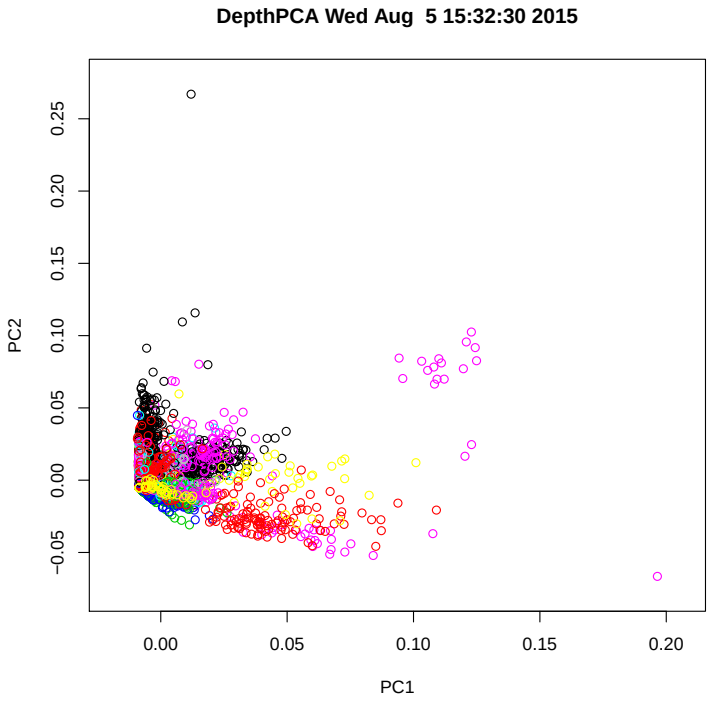


(a) Missingness PCA on a subset of UCL-ex samples.

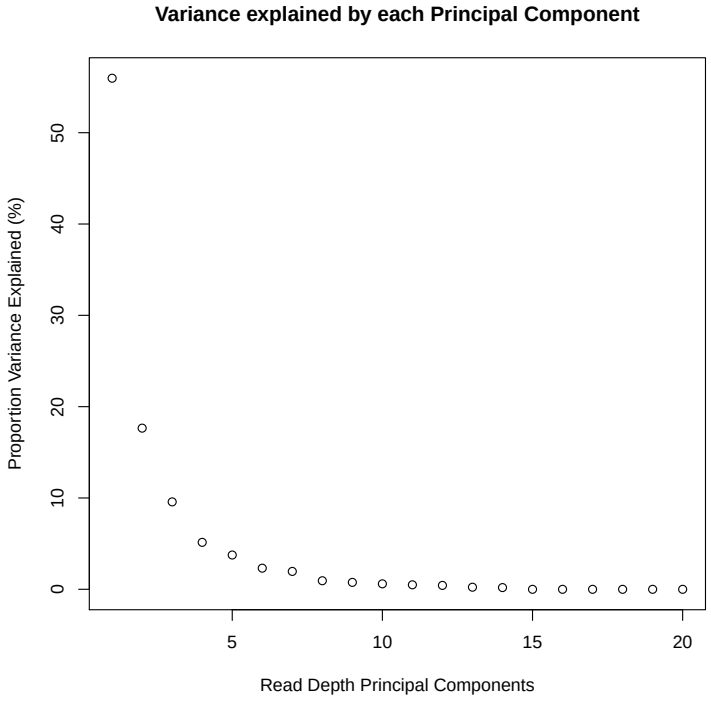


(b) Missingness and RD PCA on all UCL-ex samples.

Figure 4.16: Identifying clusters of samples based on sequencing capture technique used during preparation.



(a) Principal Component Analysis plot of the UCL-ex RD kinship matrix. The samples are coloured by research group of origin.



(b) Scree plot showing the level of variance explained by each of the top 20 PCs.

Figure 4.17: Analysis of the UCL-ex RD kinship matrix.

protective or deleterious if their frequencies in cases compared to controls vary as a result of this missingness discrepancy.

4.4.4 Identifying technical PCs that explain missingness

By taking the sum of the squares of the standard deviations of each PC, the level of variance explained by each PC was found. This showed that the first five PCs explains approximately $\sim 52\%$ (Figure 4.18A). Be that as it may, this did not offer sufficient correction when included as fixed effects in Equation 4.6.

The 1025 variants that had a pvalue that was ≥ 0.9 smaller in the Linear Regression of Equation 4.5 than the corresponding corrected pvalue in Equation 4.8 were defined as Corrected Variants (CVs). To determine if this was caused by any identifiable subset(s) of the Technical Kinship Matrix, a separate linear regression was run on each CV with one of the 1763 PC_{tech}s as covariates (1763 regressions per CV). There are 1763 PCs here because that was the number of samples we had at the time. A χ^2 test compared these models to the standard regression of phenotype on SNP with no covariates to determine which, if any, PC_{tech}s are associated with case control status ($p <= 1 \times 10^{-8}$). Figure 4.18B illustrates the extent to which different PC_{tech}s survive this threshold. This shows that even when the level of variance explained by a PC is negligible many if not all of these SNPs can be strongly associated with it. If the dataset remained small then this approach of identifying variants as artefacts by their association with PCs may be amenable. However, it is not computationally tractable at a larger scale and required too many assumptions and arbitrary filters.

4.4.5 Single Variant Model Optimisation

A cohort of 104 exomes with PID were included in UCL-ex. These samples were used as cases to refine the model. Equations 4.1 and 4.8 were used to perform a case control association on this data. 9.2% (62060 / 672504) of variants were first removed as they failed the GATK quality metric. The QQplot of the remaining variants pvalues from Equation 4.8 shows that the mere inclusion of TK did not correct the test statistics completely, as Figure 4.8A shows a deviation from the expected χ^2 distribution. Rare variants were then pruned until the distribution reflected a χ^2 distribution (Figure 4.8B-I). The threshold at which a satisfactory

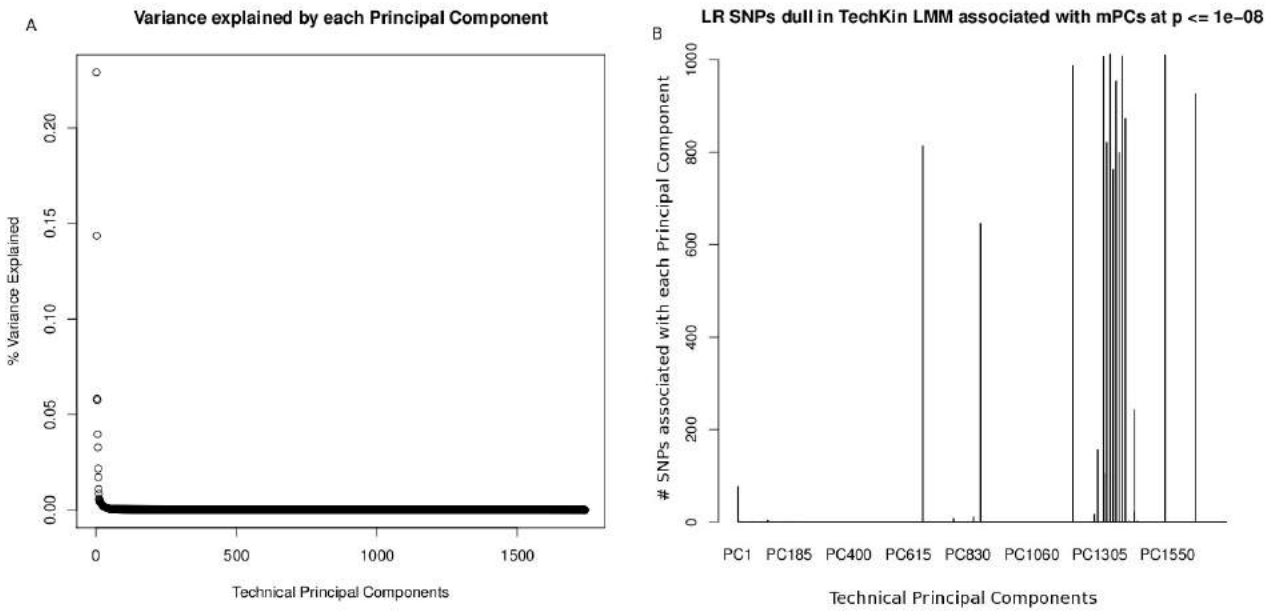


Figure 4.18: Assessing the importance of each Technical Principal Component(PC): (A) The percentage of variance explained by each technical PC. (B) The 1000 SNPs/INDELs with a pvalue difference of ≥ 0.9 between Equation 4.1s and 4.8 were tested for their associations with each of the 1763 technical Principal Components (PC_{tech}s). The x-axis displays each PC_{tech} and the y-axis is the number of these 1000 SNPs that associate with that particular PC_{tech}.

distribution was reached was assigned as variants that have ≥ 20 calls of the alternative allele.

Equations 4.1 and 4.8 were compared to a series of 20 models in which case control status was permuted randomly. This tested the theory that a large number of these technical artefacts are strongly associated with case control status, so it was expected that you would see few if any significant associations in the permutations. While such permutations would remove any true signal too, the number of variants that you would expect to be artefacts is higher than the expected number of true signals, so this approach remains valid. Figure 4.14 shows that while the permutations exhibit a much lower range of pvalues as expected, the Equation 4.1 is more extreme than 4.8, suggesting that 4.8 works to correct outliers.

A variant (E1021K) was previously identified in these samples to be a dominant gain of function that alters the *PIK3CD* gene [Angulo et al., 2013]. This was used as a positive control throughout the model development process (Table 4.7). E1021K's disease association was confirmed with an Equation 4.1 pvalue of 1.162e-23 and an Equation 4.8 pvalue of 1.809e-08. The latter being closer to the reported association of 4.767e-08.

	Type	Cohort	Variant	Nb.cases(maf)	Nb.controls(maf)	BaseP	PermutedP	CorrectedP
1	-ve Control	Retinal Dystrophy	*C1R*(c.44G>A)	3(0.038)	16(0.0025)	$7.88 * 10^{-11}$	0.88482	0.0035867
2	+ve Control	Primary ImmunoDeficiency	*PIK3CD*(c.3133G>A)	85(0.035)	2638(0)	$1 * 10^{-16}$	0.063491	$1 * 10^{-16}$

Table 4.7: Positive and negative control variants used in model development

C1R variant

The *C1R* variant from Section 4.1.1 was used as a negative control for the single variant analysis of the cone rod dystrophy cohort. The naive single variant linear regression pvalue of 1.10e-04 for chr12:7244369C>T is close to the gene based pvalue of *C1R* of 4.22e-04 from SKAT and 1.64e-04 for the binomial test. However, its artefactual status is confirmed by the corrected pvalue of 0.144 (from Equation 4.8). The Genomic Inflation Factor λ was calculated for every UCL-ex trait for three models; Equations 4.5, 4.9 and a permuted model for an idealised distribution. These λs were compared at a range of MAF thresholds (Figure 4.19).

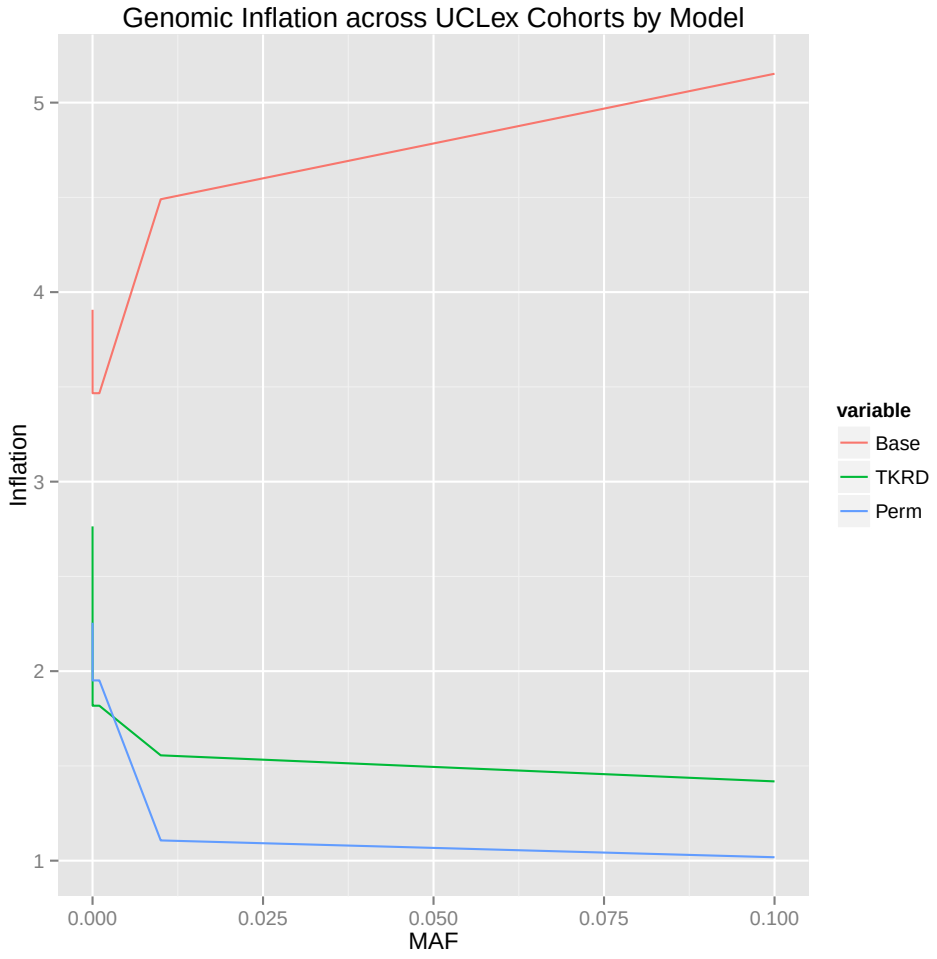


Figure 4.19: The GIF across all UCL-ex cohorts. Base (red) is the mean of the uncorrected GIFs from all cohorts, green is the corrected GIF while blue is the Permuted GIF. The X-axis indicates the MAF less than which SNPs were excluded to calculate their respective GIFs.

4.4.6 Final Single Variant Model Application

Table 4.7 in Section 4.4.5 showed that the correction applied by Equation 4.9 was performing as desired in that it successfully removed variants known to artefacts while retaining known risk loci. To further test this, it was applied to another of the UCL-ex cohorts, 800 mostly Ashkenazi Jewish Samples with Inflammatory Bowel Disease (IBD). The most associated single variants are in Table 4.8. The associated QQplots and Manhattan plots are show in Figure 4.20. While the corrected QQplot in this Figure does show a reduced Genomic Inflation, it remains quite elevated which, combined with the unusual genes with the strongest pvalues, makes it unlikely to be fully controlling for both artefacts and other noise such as PS.

rsID	SNP	Gene	Fisher	LRp	LMMp	OR	#Hom.IBD(#n799)	#Hom.ctrl(#n3535)	#Het.IBD	#Het.ctrl
rs201286142	c.1957G>A	*GRM3*	6.16E-04	4.78E-02	2.30E-16	25	0	0	3	5
rs184616940	c.2624C>T	*LRRCC1*	2.34E-18	4.59E-11	1.21E-15	90	0	0	14	6
rs77786095	c.1376G>T	*DTX3L*	1.08E-07	3.49E-07	2.09E-15	44	0	0	6	6
rs200843707	c.658G>A	*GUF1*	4.39E-17	1.75E-12	4.39E-15	60	0	0	14	9
rs201337101	:c.2597T>C	*ITGAM*	1.77E-05	3.77E-14	6.23E-15	21	0	1	5	8
rs139134493	c.2328A>G	*TTC27*	1.99E-11	6.68E-07	1.41E-14	55	0	0	9	7
rs139555612	c.1966T>A	*RTN4*	6.66E-04	≤1.00E-16	1.79E-14	25	0	0	3	5
rs104895423	c.662T>G	*NOD2*	5.53E-05	1.27E-03	1.60E-07	15	0	0	5	15

Table 4.8: IBD Single Variant Test Results, with 799 cases and 3535 controls. SNP details the position of the tested variant (hg19). Gene is the HUGO name for the gene in which the SNP resides. Fisher is the pvalue from Fisher's exact test. LRp is the Linear Regression pvalue with no covariates or kinship matrices. LMMp is the pvalue from Equation 4.9. OR is the risk odds ratio. 'Homs' are homozygotes for the minor allele, while 'Hets' are heterozygotes.

4.4.7 Gene Based Model Optimisation

Variants that were not deemed to be artefacts based on the linear regression of Equation 4.12 were filtered further by selecting for rare variants and those that were either non-synonymous, LOF or splicing. The region based Binomial and SKAT tests were then run on these variants for each gene separately for the PID, ARVC and SCD cohorts. The top PID gene for the Binomial test was *CASC5* (p <5.765e-06). The counts for the 5 top genes for the PID cohort are in (Table 4.9) and the QQplots for both tests of all three cohorts are in Figure 4.21 for the Binomial pvalues and Figure 4.13 for SKAT. As these figures show, this gene based model did not work so no reliable results were generated.

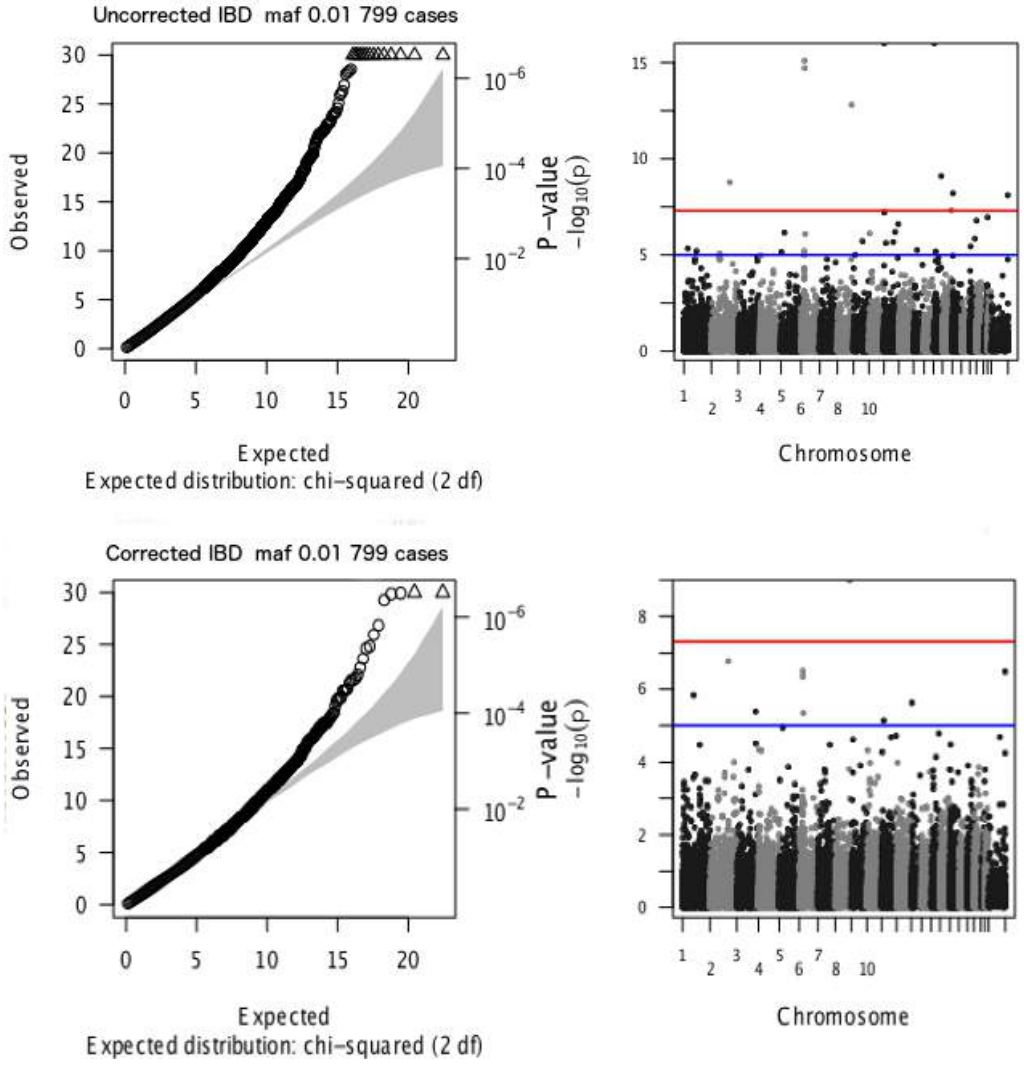


Figure 4.20: Artefact correction in the Inflammatory Bowel Disease cohort of UCL-ex. The qqplot and the Manhattan plot in the top row show the uncorrected LMM results. In contrast, the bottom row displays the results from Equation 4.9.

4.4.8 REML Estimates of variance explained by Kinship component

The estimations of variance explained by the TK/RD Kinship matrices initially showed that some phenotypes were very highly correlated with these kinships, indicative of being artefacts. To start, the only variant filtering performed was removing variants that GATK 'PASS' flag. However, as Figure 4.22 shows, the GATK flag was rather uninformative. This meant that many cohorts had $\geq$ 95% variance explained by

Gene	Position	Case Counts(n=143)	Control Counts(n=1,956)	SKAT	Binomial
CASC5	chr15:40895128-40954311	14	31	1.245890e-05	5.765778e-06
LRRC46	chr17:45909365-45914403	9	12	4.838810e-05	1.451893e-05
C4orf17	chr4:100434240-100463258	12	25	3.836317e-04	1.657423e-05
PGAM2	chr7:44102399-44105115	7	6	2.188516e-04	2.137651e-05
PZP	chr12:9302172-9360933	11	22	1.189751e-03	2.781703e-05

Table 4.9: Top 5 PID candidate genes based on the binomial test. The criteria for retaining variants are: GATK Variant Quality score of PASS, MAF (MAF of $\leq 0.3\%$) , $\leq 10\%$ missingness across all samples, non-synonymous, LOF or affecting splicing.

the kinships. This was thought to explain why the initial corrections appeared to work well; it was in fact explaining most of the variance in the data which lead to an inability to detect any true signal.

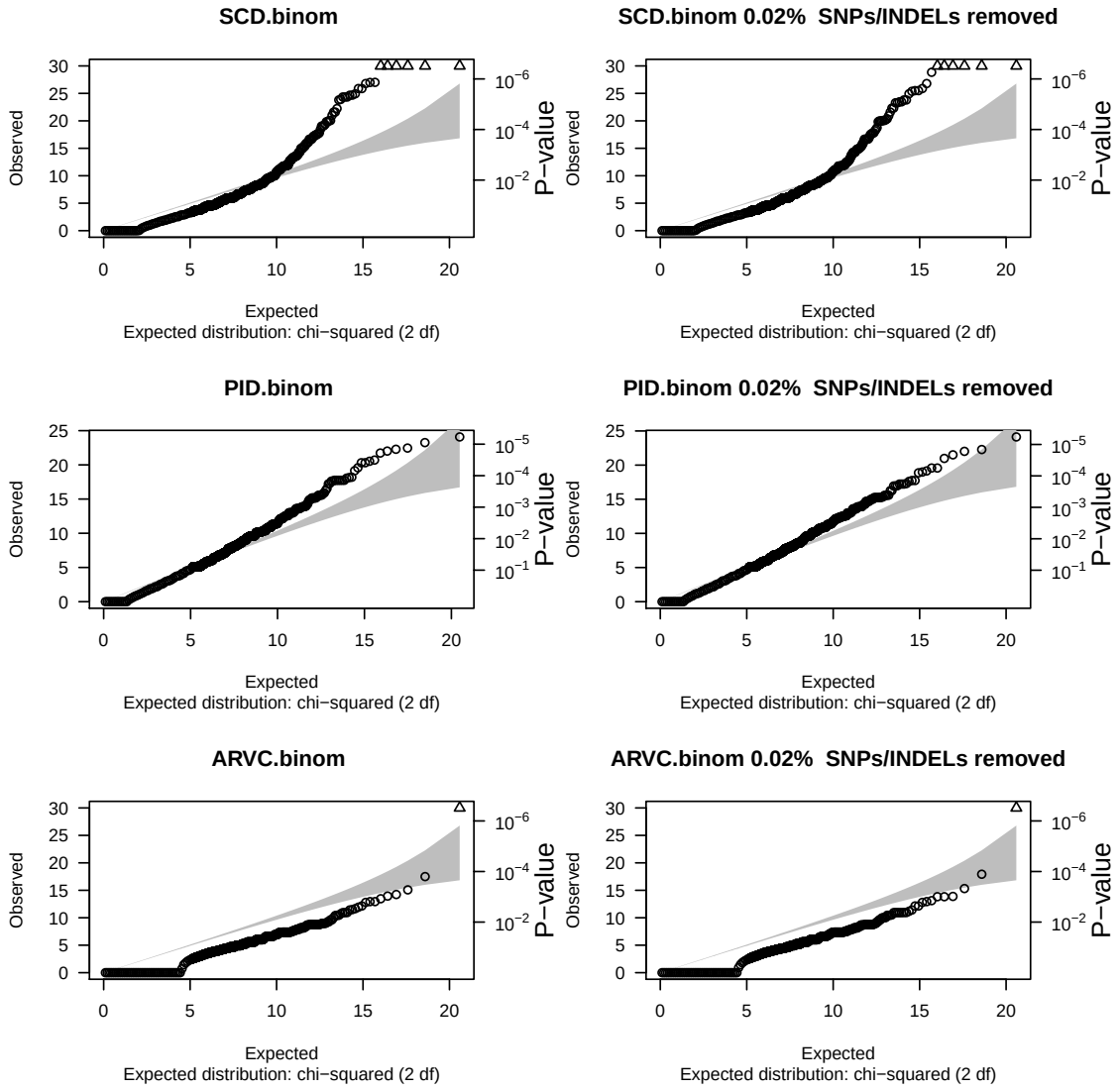


Figure 4.21: Binomial Gene based tests for the PID, ARVC and SCD cohorts. Each circle is the gene based pvalue from the binomial test. QQplots compare the observed distribution of pvalues to the expected Chi Squared distribution. Before the pvalue for each gene is calculated, some variants are filtered/removed. On the left graphs, the criteria for retaining variants are: GATK Variant Quality score of PASS, MAF (MAF of $\leq 0.3\%$) , $\leq 10\%$ missingness across all samples, non-synonymous, LOF or affecting splicing. (B) For the graphs on the right, the same criteria are used but additionally variants are filtered based on the technical PCA. The first ten principal components (PC) are included in the linear regression as covariates. SNPs that are associated with the technical PC are removed. The percentage of SNPs/INDELs removed across all genes is included in the figure titles.

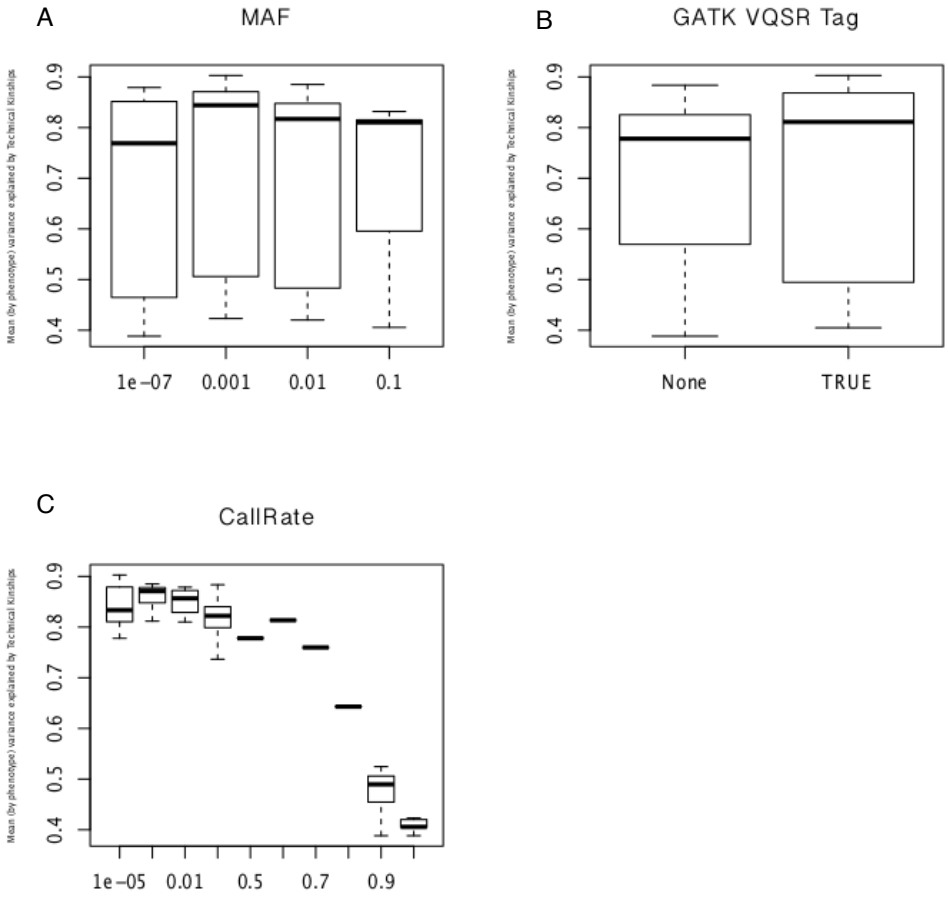


Figure 4.22: Effect of SNP filtering using different criteria on amount of variance explained by technical kinship. (A) Filtering SNPs by MAF. The X axis indicates the minimum MAF of the retained SNPs and the Y axis shows the averaged level of variance explained by the Technical Kinship for all UCL-ex phenotypes. (B) Low quality SNPs that did not receive a 'PASS' flag from the GATK VQSR test were removed. (C) The required variant call rate is increased across the X axis.

4.5 Arrhythmogenic Right Ventricular Cardiomyopathy

4.5.1 ARVC-UCLex Single Variant Association Tests

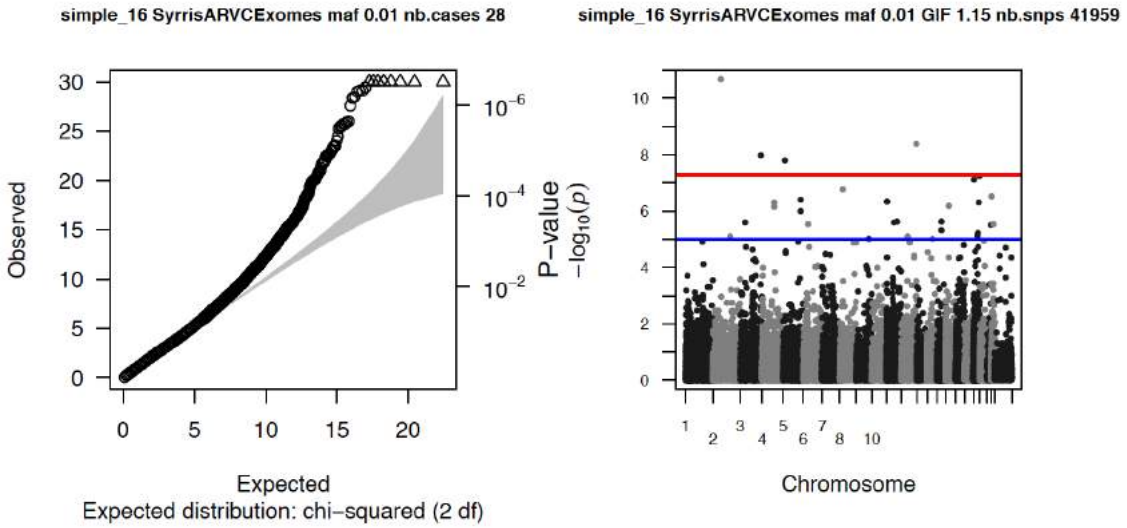
28 ARVC exomes were also included in the UCL-ex consortium. They were subjected to the Single Variant analysis (Equation 4.9) of common variants and Gene based testing for rare variants as previously described. Figure 4.23 illustrates the QQ plots of both the uncorrected and corrected models. The 5 SNPs most strongly associated with ARVC are in Table 4.10. These results should be interpreted with caution however as 28 samples does not provide adequate power to detect all but the largest signals and may yield many artefacts. The gene based testing procedures involving SKAT and the binomial test were additionally applied to the ARVC cohort, with Table 4.11 detailing the genes with the strongest association.

rsID	SNP	Gene	Fisher	LRp	LMMp	OR	#Hom.ARVC(#n28)	#Hom.ctrl(#n4306)	#Het.ARVC	#Het.ctrl
rs368209124	c.1162-3C>T	*COL9A2*	$1.59 * 10^{-4}$	$\leq 1. * 10^{-16}$	$\leq 1. * 10^{-16}$	172	0	0	2	3
rs149175095	c.1123A>C	*PHF7*	$3.14 * 10^{-4}$	$\leq 1. * 10^{-16}$	$\leq 1. * 10^{-16}$	106	0	0	2	5
rs199640194	c.20318G>A	*TTN*	0.034	$3.68 * 10^{-9}$	$\leq 1. * 10^{-16}$	0	1	0	0	4
rs150671437	c.1619G>A	*EFHB*	$9.46 * 10^{-4}$	$\leq 1. * 10^{-16}$	$3.42 * 10^{-10}$	54	0	0	2	9

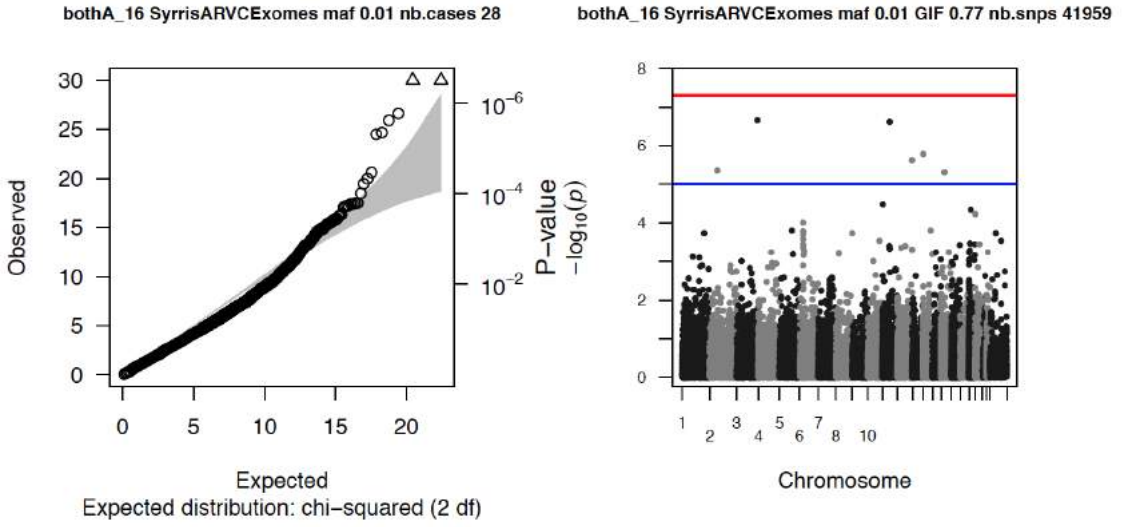
Table 4.10: ARVC Single Variant Results. SNP details the position of the tested variant (hg19). Gene is the HUGO name for the gene in which the SNP resides. Fisher is the pvalue from Fisher's exact test. LRp is the Linear Regression pvalue with no covariates or kinship matrices. LMMp is the pvalue from Equation 4.9. OR is the risk odds ratio. 'Homs' are homozygotes for the minor allele, while 'Hets' are heterozygotes.

Gene	Position	Case Counts(n=16)	Control Counts (n=4318)	SKAT	Binomial
TAS2R40	chr7:142919173-142920122	6	7	$1.83 * 10^{-3}$	$6.35 * 10^{-9}$
ANO5	chr11:22225349-22301267	5	34	$2.80 * 10^{-4}$	$1.27 * 10^{-4}$
PPP1R3F	chr23:49126534-49143288	2	1	$1.42 * 10^{-3}$	$4.72 * 10^{-4}$
CIITA	chr16:10989219-11017124	4	31	$1.09 * 10^{-2}$	$9.65 * 10^{-4}$
ATF7IP2	chr16:10524564-10576102	3	13	$4.20 * 10^{-3}$	$9.90 * 10^{-4}$

Table 4.11: Top 5 ARVC candidate genes based on the binomial test using the rest of UCL-ex as controls. The criteria for retaining variants are: GATK Variant Quality score of PASS, MAF (MAF of $\leq 0.3\%$) , $\leq$ 10% missingness across all samples, non-synonymous, LOF or affecting splicing.



(a) Left - qqplot of the uncorrected ARVC analysis for variants with a MAF of $\geq$ 1%. Right - Manhattan plot showing the associations per chromosome. Red horizontal line is pvalue of $1 * 10^{-8}$ and blue is $1 * 10^{-5}$.



(b) Left - qqplot of the corrected ARVC analysis for variants with a MAF of $\geq$ 1%. Right - Manhattan plot showing the associations per chromosome. Red horizontal line is pvalue of $1 * 10^{-8}$ and blue is $1 * 10^{-5}$.

Figure 4.23: Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) single variant mixed model association results.

4.6 Comparing Coding and NonCoding variants in ARVC to HCM

In addition to the main UCL-ex work, this analysis was applied to 359 Arrhythmogenic right ventricular cardiomyopathy (ARVC) and 875 Hypertrophic Cardiomyopathy (HCM) genome/exome samples. These samples were prepared in the same lab, using the same capture methodology. From HCM "plate 3 rerun" on, the sequencing platform differed - it occurred on a HiSeq2000 instead of a GAIIx, with increased multiplexing for plates 4 and 5 (96 samples) and again for 6 (120) and 7,8,9 (128). The exact sample breakdown by platform is GaIIx (252 samples), HiSeq2000 (12) and HiSeq2000.Multiplexed (695). A Principal Component Analysis was performed on the missing/nonMissing genotype matrix to ascertain what the missingness patterns in the data were. The initial step in analysing this consisted of determining if there was a significant difference that correlated with phenotype. Figure 4.24A reveals that these technological disparities affect the PC_{tech} more than the disease differences.

Processing this many samples is routinely performed in batches. These samples were prepared in 12 distinct batches, some of which involved re-running samples for Quality Control purposes. Figure 4.24B shows that this batch effect is readily visible. While some batches are distinct from each other with almost no overlap, the majority are similar. This effect is less powerful than that influenced by the transition from the Illumina GAIIx to its HiSeq 2000 sequencing system (Figure 4.24C).

As these samples included non coding regions, a gene based test was infeasible. While a region can be defined in any arbitrary way to offer an alternative, this has not yet been performed. As noted previously, TK is better suited to correcting common artefacts than rare ones. Rare variants, those with less than 20 calls of the non-reference allele between cases and controls were therefore excluded. Any related samples were removed before the analysis as part of the standard QC based on clinical pedigree data and plink estimates of relatedness. The top 5 variants are listed in Table 4.12 and the number of variants that are significant at a range of levels summarised in Table 4.13.

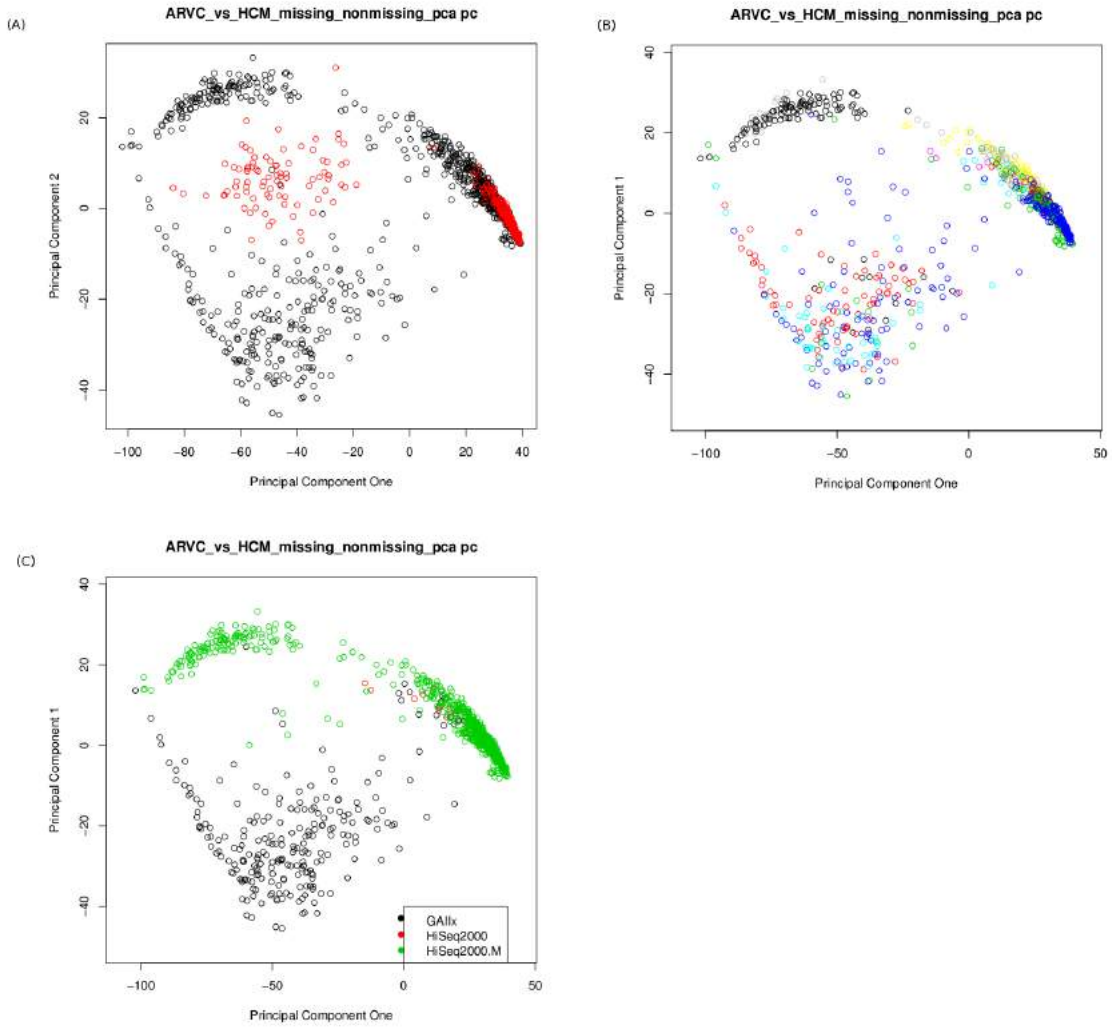


Figure 4.24: Technical PCA of ARVC vs HCM: (A) Arrhythmogenic right ventricular cardiomyopathy (ARVC) in red compared to Hypertrophic Cardiomyopathy (blue). (B) Samples were prepared in batches. (C) Different sequencing technologies.

4.6.1 ARVC, HCM and UCLex joint artefact analysis

As discussed already in this thesis, pre-sequencing combination of cohorts can lead to technical artefacts. Figure 4.25 displays such an effect when one integrates the ARVC,HCM and UCL-ex cohorts, retaining the variants that are called in both cases and controls. Additionally, the technical PCA was performed on the rotated matrix. Figure 4.26 shows the top two PCs from this. Figure 4.26B shows these loadings coloured

Variant	Gene	ARVC Counts(n=407)	HCM Counts(n=957)	ARVC.Freq	HCM.Freq	P_T
rs193922652	*MYH6*	12	35	3.34%	4%	$5.305 * 10^{-41}$
rs111679193	*PKP2*	30	52	8.36%	5.94%	$1.302 * 10^{-14}$
chr17:68174142_-_T	*KCNJ2*	3	28	1.11%	3.2%	$2.260 * 10^{-14}$
chr2:179660461_T_-	*TTN*	70	149	19.5%	17.03%	$4.029 * 10^{-14}$
chr18:28681054_-_G	*DSC2*	20	58	5.57%	6.63%	$1.016 * 10^{-12}$

Table 4.12: The most significant common ($\geq$ 20 total calls) SNPs/INDELs from the Linear Mixed Model with technical Kinship correction of the ARVC vs HCM comparison. The columns, in order from left to right, are the genes containing the variant, its exact position, the number of counts in the ARVC and HCM samples, the resultant frequencies and corrected pvalue. The absolute genomic position is reported for the variants that do not effect the coding sequence.

Threshold	Nb.SNPs.Below	Nb.SNPs.Above
1e-03	39	10068
1e-04	25	10082
1e-05	19	10088
1e-06	12	10095
1e-07	12	10095
1e-08	10	10097
1e-09	9	10098
1e-10	8	10099
1e-11	7	10100
1e-12	4	10103
1e-13	4	10103
1e-14	1	10106

Table 4.13: The number of SNPs/INDELs that are significant in the ARVC/HCM comparison at a number of thresholds.

by gene of origin. Gene was used to represent genome location and from this you can see that missingness varies systemically across the genome. This graph is dominated in the centre by Titin. Given that Titin dwarfs the other genes in length, at some 34Mb long, it is expected to contribute the most to the PCA.

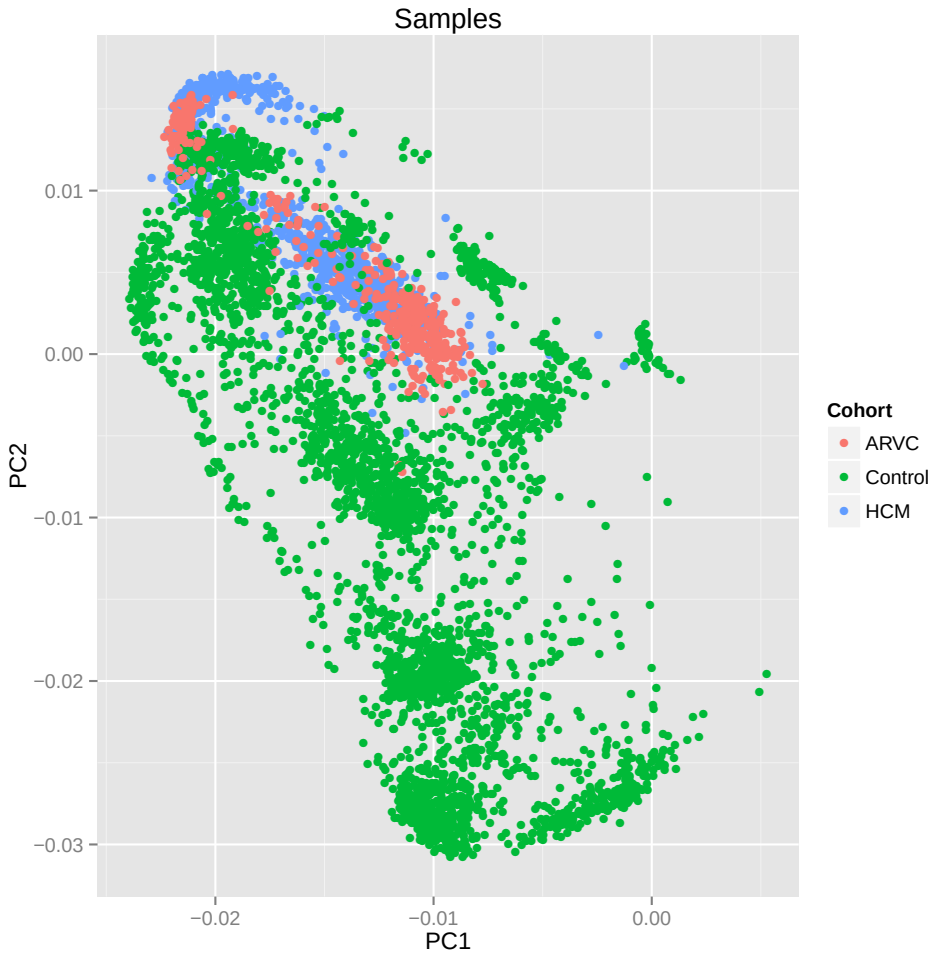


Figure 4.25: Technical PCA of the HCM/ARVC Joint Analysis. The ARVC (red), HCM (blue) and UCL-ex control (green) samples are shown.

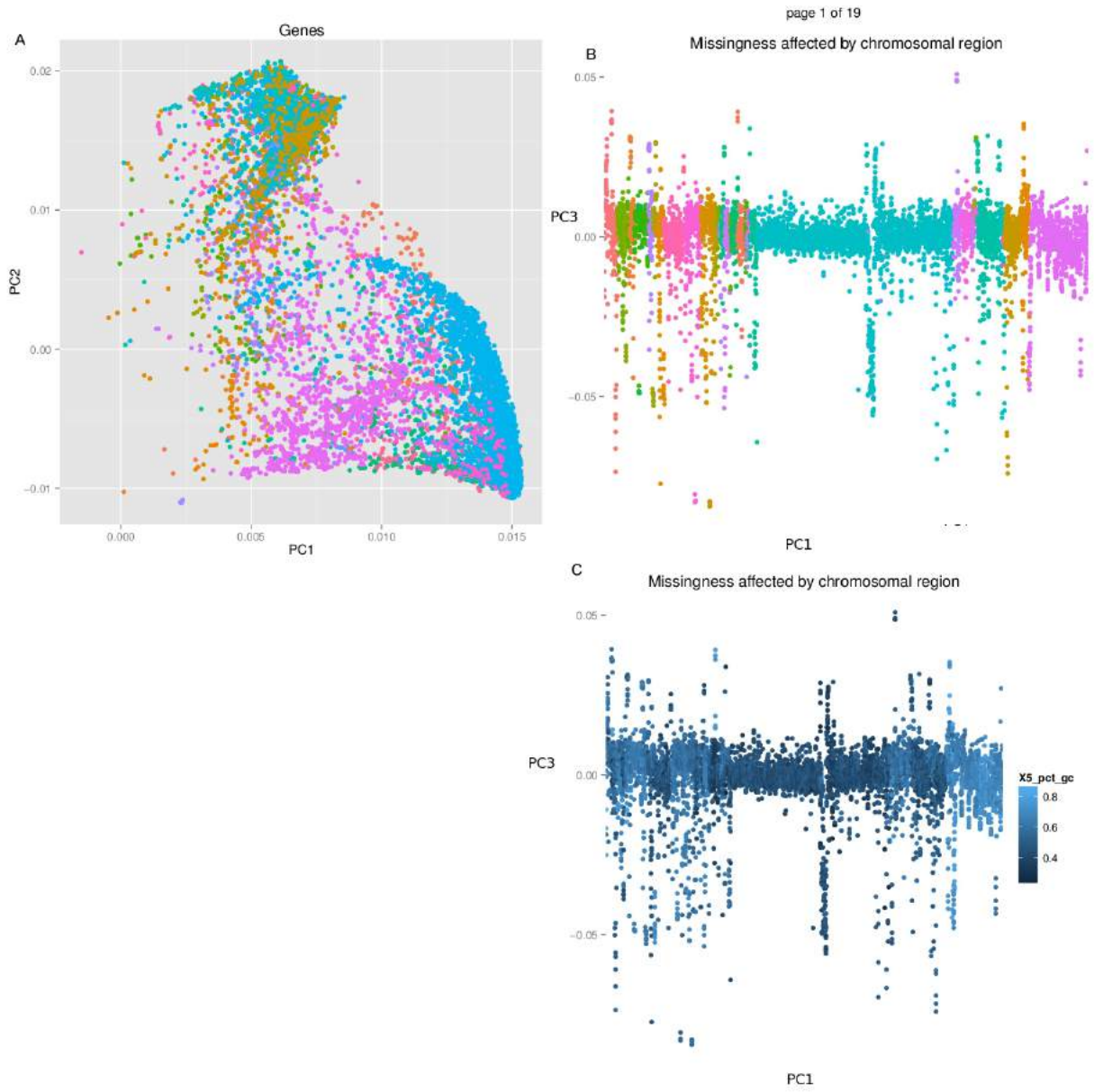


Figure 4.26: Technical PCA of the ARVC/HCM joint analysis. Here, the data is rotated so that the Eigenvectors correspond to the variants, rather than individuals, which is the norm. (A) PC one (X-Axis) against PC2. The points represent SNPs and INDELs and are coloured according to gene of origin. (B) PC1 against PC3 of the same data, again coloured by gene. (C) PC1 against PC3 but here coloured based on GC content of SNPs, as defined by ± 50 base pair bins around each variant.

4.7 Discussion

4.7.1 Single variant model optimisation

Studies that perform case control associations through the analysis of exome sequencing data have a variety of sample preparation related confounders to differentiate from true signals [Nothnagel et al., 2011; Lam et al., 2012]. We have herein described a model that implements a novel approach to deal with such cryptic artefacts. A kinship matrix can be calculated that estimates the extent to which pairwise similarity between individuals is based on the missing/nonMissing status of their respective variants, Single Nucleotide Polymorphism (SNP) and/or Insertions/Deletions (INDELs). By including this alongside a RD kinship matrix in a Linear Mixed Model that tests for an association between disease status and genotype, you can get a measure of association that is free from noise caused by SNPs with spurious call patterns.

For the association tests performed here, an additive genetic model was assumed. This is the norm in GWAS and operates by representing the major (more common) allele as 0 and the minor allele as 1. Homozygote wildtypes are therefore given a count of 0 for a given SNP, while heterozygotes are 1 and homozgygotes for the minor allele are 2, respectively. The 4334 samples in the UCL-ex consortium that served as the test dataset for model development were exome sequenced, which generated 900,000 calls of SNPs or INDELs.

Long a mainstay of genetics, linear regressions of all single genotyped SNPs has been robustly studied in association studies [Lourenço et al., 2011]. Applying linear regression, while controlling for population stratification, to a case control analysis of all groups in UCL-ex yields an inflated false positive across many SNPs (Figure 4.9). This makes it difficult, if not impossible, to gain an accurate idea of what the true disease causing/associated variants are. To improve the ability of the model to correct for the data artefacts driving this Type 1 error inflation, a Principal Component Analysis was performed on the binary missing/nonMissing genotype matrix (PC_{tech}). The Principal Component plot in Figure 4.4 reinforces the notion that factors such as sample preparation or sequencing chemistry used to process samples can influence the variant call rate more than the sample phenotype. An effective technique to overcome such noise will be a boon to

modern exome sequencing studies that require ever larger sample sizes. This will instantly make available many samples, that until then will have been incompatible because of such technical artefacts. The Technical Principal Components were included in two forms of linear regression, which differed based on whether the phenotype or the genotype was the dependent variable. For the former, a corrected pvalue was obtained for all SNPs and CNVs. This correction failed to effectively correct the data to a χ^2 distribution (Figure 4.9). This was the case for the three cohorts tested even when the analysis was restricted to solely the common variants.

To gain more power than the process of including an arbitrary number of PC_{tech}s in the association tests, a linear mixed model (LMM) was then used instead of a linear regression. LMMs contain a random effects component, which can include a kinship matrix that traditionally is a measure of the pairwise genetic similarity between all individuals in the study. We have modified this to what we call a "Technical Kinship" matrix (TK). The kinship is calculated on the missing/nonMissing genotype matrix. This estimates the extent to which observations' genotyping success rates are similar. Figure 4.8 shows that the mere inclusion of this matrix does not yield the expected χ^2 distribution. Biases caused by spurious calls from different capture technologies or sequencing platforms are unlikely to be overly common [Nothnagel et al., 2011]. Through a process of repeated pruning, it was found that by restricting the analysis to variants that had at least 20 calls of the alternative allele, Equation 4.8 adequately adjusts for artefacts (Figure 4.8) for some of the traits tested.

Not all artefacts will manifest as a binary missing/nonMissing factor. As shown in Figure 4.17, a PCA of a RD kinship matrix does cluster samples based on research group of origin. Figure 4.16B includes the sequencing platform and chemistry information, which is known for just a subset of samples within UCL-ex. The fact that this is captured by using RD as a proxy as per Figure 4.17 is thus clear. This was the motivation behind adapting Equation 4.8 to include a RD kinship matrix, yielding Equation 4.9, the final model used for single variant testing. Application of this model gave a pvalue of 0.00358 for the negative control and $1 * 10^{-16}$ for the positive control in Table 4.7, which shows the model has sufficient sensitivity and specificity.

4.7.2 Model application to Crohn's Disease

Linkage analysis was employed to identify the first CD susceptibility locus [Hugot et al., 1996]. 25 Caucasian families, each with at least two affected individuals were genotyped for 270 polymorphic markers that were spread throughout most of the genome. A region on chromosome 16 was identified as conferring susceptibility to CD ($p = 0.01$). In-depth analysis of this locus subsequently identified variations in the *NOD2* gene as involved in CD susceptibility. Two missense variations and a frame-shift altered the leucine rich repeat domain (LRR) in *NOD2* [Hugot et al., 2001]. Via the LRR, *NOD2* detects the presence of muramyl dipeptide, a component of bacterial cell walls. It also activates nuclear factor Kappa B (NFB). *NFB* is a major transcription factor that is involved in cancer, inflammation, immunity and synaptic plasticity [Gilmore, 2006; Ogura et al., 2001]. As a result of recent GWAS meta-analyses, there are now to thought to be at least 71 CD susceptibility loci [Franke et al., 2010].

Equation 4.9 was used for single variant testing on the IBD cohort in UCL-ex (Table 4.8). The *NOD2* SNP Chr16:50744565T-G has a pvalue of $1.60 * 10^{-7}$. While not the strongest association for this trait, which is for a variant in *GRM3*, a gene with no known association with Crohn's Disease, it remains evidence of model efficacy. In addition, a variant in *ITGAM*, 16_31336912_T_C, has a pvalue of $6.23 * 10^{-15}$. *ITGAM* one has had a disputed role in Crohn's for some time [Kenny et al., 2012; de Jong et al., 2003]. It may related to the functioning of a microRNA hsa-miR-155, reduced levels of which have been shown to have a protective effect against colitis while lowering the number of CD11b+ T helper cells [Singh et al., 2014]. The QQ-plot in Figure 4.20 still displays evidence of inflation which explains the domination of the list by presumed false positives. Despite best efforts, it was not impossible to improve this further. The difficulty of interpreting this cohort was exacerbated by the fact that 203/799 (25%) of the cases came from 2 large families. While in the case of the SCD analysis it was practical to remove related individuals without having too deleterious an impact on sample size, here that is not the case. This *NOD2* variant was seen in both families and in unrelated cases however so it remains a plausible result.

4.7.3 Model application to SCD and ARVC

68 patients were diagnosed with SCD and their exomes included in UCL-ex. A basic linear regression was initially used. Many variants were flagged as highly significant but were based upon the presence of a singleton, typically because of one call of the non-reference allele across all samples. This can be rectified by removing variants with a MAF of <1:1000 or by using an exact test.

The SCD samples were initially analysed with the non-proband family members excluded. Table 4.3 contains those that have a pvalue of $\leq 1 * 10^{-10}$. The QQplots in Figure 4.12 show that this data has a high inflation factor meaning it is not clean enough to be interpretable and informative clinically. The most significant variants are in open reading frames so are most likely not transcribed.

Table 4.4 contains the 9 variants that have a pvalue of $\leq 1 * 10^{-10}$ for the SCD cohort when the J wave family is included. Of these, three are in genes that have previously been shown to be associated with heart development or disease. *ADAM19* has been shown to have a role in the development of the endocardial cushion and congenital heart disease [Kurohara et al., 2004; Goldmuntz et al., 2011]. Just last year, *FSTL1* was found to a potent activator of regeneration of the adult mammalian heart following myocardial infarction [Wei et al., 2015]. Finally, the expression of *BTG2* is increased when oxidative stress occurs in cardiomyocytes [Choi et al., 2013]. However, it does have many other functions so is likely a false positive [Tong et al., 2015]. None of these genes are already reported strong candidates for SCD so caution must be paid to their veracity. All variants in this list, excluding chr5:156915410C-T and chr9:20944681C-G, are present only in members of a single family pedigree (Figure 4.11). This family was identified after the proband presented with Ventricular Fibrillation, and was subsequently identified to have a J wave abnormality on ECG, along with four immediate family members. This includes both parents and 2/6 siblings, indicative of a recessive model of inheritance. None of these variants co-segregated with the J wave phenotype in this family, so are unlikely to be causative and are therefore thought to be benign private mutations.

Validation of private mutations can often be complicated by the large number of rare variants with uncertain effects that are present in the general population, even in candidate genes. Furthermore, given

the incomplete penetrance and variable phenotype of SCD, this approach would be even more difficult here. Efforts to improve the predictive accuracy of variant pathogenicity is an ongoing effort. This includes improving algorithms to predict biological function or evolutionary conservation (eg SIFT and PolyPhen) and more clinically oriented work that attempts to associate certain variants or genes with specific presentations of a given condition [Lopes et al., 2014; Syrris et al., 2007]. It is expected that in the future such progress will help with the interpretation of variant lists generated by HTS such as that discussed here. As it stands, including related individuals in the study complicates the interpretation while offering little benefit which is in agreement with most of the literature that espouses removing them or controlling for relatedness.

The ARVC analysis' top SNPs are in Table 4.10. Only four reach a pvalue of $\leq 1 * 10^{-10}$. Three are in genes with no known association with ARVC. The fourth however, the non-synonymous chr2:179482565C-T rs199640194 is in Titin exon 253. This however is a singleton with an exact Fisher pvalue of 0.034. As stated already, only 28 cases were available for this analysis. This does not have enough power to detect anything except the strongest of signals so a lack of strongly significant pvalues was to be expected.

4.7.4 Comparing the genetic architecture of ARVC to that of HCM

The clinical presentations of HCM and ARVC is somewhat similar. By comparing ARVC samples directly against HCM samples, it was thought that insight may be gained about any genetic variants that are more associated with one cardiomyopathy than the other. A targeted sequencing approach including the exonic and flanking/intronic noncoding regions of 73 genes was sequenced. Table 4.12 details the 5 variants most able to discriminate between ARVC and HCM pathogenesis. With a pvalue of 4.447e-33 for the chr14:23858281GC_- variant, the *MYH6* gene seems more strongly associated with HCM than ARVC. This is in line with previous studies, one of which recently showed that allele specific silencing of certain *MYH6* transcripts suppresses HCM [Carniel et al., 2005; Jiang et al., 2013].

PKP2 variants have been robustly shown to cause ARVC [Li Mura et al., 2013; Roberts et al., 2013; Cerrone and Delmar, 2014]. A recent study of 90 subjects identified 78 variants in known ARVC genes; *PKP2* mutations consisted 31 (58%) of these. Furthermore, *PKP2* carriers were significantly more likely to

exhibit Ventricular Tachycardia than those with other putatively causative variants [Bao et al., 2013]. This lends credence to our findings that the chr12:33026249T>C SNP in *PKP2* is significantly more common in ARVC cases than HCM cases (8.36% & 5.94%, respectively, p = 1.302958e-14).

4.7.5 Gene based tests

Association tests have limited power to detect rare variants at the single variant level. It could therefore be argued that the inability of Equation 4.8 to correct artefacts when their MAF is low is of little consequence. This has led to a plethora of solutions that involve the combination of rare variants into region based testing procedures. These can be broadly categorised as the Cohort Allelic Sum tests [Morgenthaler and Thilly, 2007], the Combined Multivariate and Collapsing Tests [Li and Leal, 2008a], Weighted Sum Tests [Madsen and Browning, 2009] or Kernel Association tests. A multivariate test such as the Sequence Kernel Association Test [Wu et al., 2011] combines single variant test statistics while not declaring alleles that are more frequent in cases as necessarily deleterious, as is the case with some alternative methods. As Equation 4.8 controlled for artefacts in common variants but not rare variants, SKAT was applied to UCL-ex. Variants were excluded from the SKAT procedure if an initial linear regression, where genotype was the outcome, indicated that the top ten technical Principal Components indicated artefactual status. As described in Section 4.2.8, SKAT and a binomial test were applied to the rare variants that were predicted to be of functional import (those that cause non-synonymous, LOF or splicing changes). By comparing the gene based tests with no correction, as shown in Figure 4.21, to the scores from the corrected model (Figure 4.13), it appears that the correction has no noticeable impact. This is similar to the results from Equation 4.6 that showed that the first ten technical principal components fail to adequately control for artefacts.

4.7.6 Application to SCD

A different gene based testing procedure was used that compared the linear mixed model with technical kinship (TK) correction to permutations that established a null distribution for the data free from phenotype specific batch effects. Here, the X-chromosome gene *SPACA5B* is strongly associated with SCD. Without

TK, *SPACA5B* has a pvalue of 1, but its inclusion changes that to 1.26e-08, two orders of magnitude lower than the lowest permuted pvalue. Whilst this is encouraging evidence for an association, knowledge of a biological relevance is lacking. It is known to be expressed in the acrosome, the cap like structure of the sperm that is involved in fertilisation (GeneCards). It may yet have a functional role in SCD. The fact that it is on the X chromosome may reflect the epidemiological observation that SCD has a higher prevalence in men than women.

It is not uncommon for association studies to flag genes that have no readily relevant role that could be linked to disease pathogenesis. Typically, such findings are tested by examining more cases and controls, through sequencing or from publically available data, to identify if the association remains. This may then be combined with functional examination of the protein from cell culture to a model organism such as a mouse. While this approach can work well, it can be onerous. To increase confidence in an association, we recommend subjecting significant genes to an additional round of (eg 10000) permutations. If the corrected pvalue remains outside the range of all of the permutations, then it is likely a real association.

Finding a novel causative gene in this way is unlikely. Through using this correction, we expect to render artefacts insignificant and not necessarily to increase the p-value of true associations. This hypothesis combined with the paucity of supporting literature lead me to consider genes such as SPACA58 to be false positives.

4.7.7 Chapter summary

This chapter is the first demonstration of an attempt to use alternative Kinship matrices to control for factors other that PS. While the final model used does show an improvement in the distribution of association test statistics across a range of cohorts, clear inflation of the test statistic distribution remains. This is in addition to the possibility that real signals may also be removed by the correction. It was not possible to find a single kinship or particular combination thereof that worked sufficiently well across all cohorts. Thus, the major limitation is the inability to distinguish technical artefacts from actual association signals.

Chapter 5

Discussion

5.1 Application to other non-cardiovascular cohorts

5.1.1 Single variant analysis

Model 4.9 was applied to the principal cohorts within UCL-ex. This included SCD (Section 4.3.2), IBD (Section 4.4.6) and ARVC (Section 4.5.1) cohorts. The results for 12 additional phenotypes are included in the Appendix but not discussed further. These include cohorts with:

- Huntington's Disease
- 3 Ophthalmology conditions
- IBD - An additional ethnically distinct Icelandic IBD cohort
- An unknown Neurological condition
- A Dermatology condition
- Keratoconus
- Primary Immuno Deficiency

- A Prion condition
- A Mitochondrial defect
- Bone Marrow Failure

Two of these, the IBD and ARVC cohorts, proved quite amenable to this model, with principal genes NOD2 and TTN remaining significant with this correction. An in-depth clinical knowledge of these conditions is required for a full evaluation of the veracity of any single variant or gene lists. In general however, this combined with the improved genomic inflation factor (Figure 4.19) indicates that the model works well in at least some situations.

5.1.2 Dealing with rare variants or limited cases

As mentioned already, various methods for accurate rare variant association tests were attempted. Published methods such as SKAT generally perform better than basic binomial or Fisher tests because they are more capable of modelling complex genetic architecture. The gene based results for ARVC and HCM in Section 2.2.2 robustly agreeing with the known genetic architecture of these conditions reinforces this point. Even these methods struggle however when faced with factors such as cryptic relatedness or batch effects that cannot be readily controlled for by currently available methods. That is to say, a mixed model methodology for region based testing that can incorporate multiple kinship matrices and fixed covariates such as that developed here for single variants has not yet been developed. The attempts discussed here are a start but it is beyond the scope of this PhD to progress it further. As a result, until this is robustly developed, any results from these models under development should be viewed with caution.

In order to refine our understanding of the clinical sub-types of conditions such as HCM and ARVC, more specific phenotypes should be used in association studies. For example, patients with apical HCM have been shown to have less fibrosis and diastolic dysfunction than those without apical involvement [Kim et al., 2015]. This difference may have a genetic basis and could perhaps be tested by separating cases into apical and non apical. The obvious problem with this approach would be the resultant reduction in sample

size lowering the power to detect association. Table 4.10 shows the association results for ARVC based on 4334 samples. It also highlights the difficulty faced when even a number as large as this contains just 28 cases. It was impossible to get reliable results with such few cases so a balance should be achieved between statistical power and clinical utility. Some methods have been developed in an attempt to improve power amidst phenotypic heterogeneity. 2010 saw the discussion of a multinomial regression modeling framework that categorised type 2 diabetes(T2D) cases by Body Mass Index. This allowed discrimination to be made between the genetic basis of obese and non-obese forms of T2D [Morris et al., 2010]. This was improved upon by the development of a multiclass likelihood ratio approach which determines itself the optimum number of subphenotypes and builds a risk model prediction for each [Wen and Lu, 2013].

5.2 Limitations of the methodology for technical artefacts

My results show an improvement in the distribution of association test statistics across a range of cohorts. The advantages to including alternative kinship matrices, such as Read Depth and Missingness have been demonstrated. However, for several of the cohorts considered in this thesis, clear inflation of the test statistic distribution remains. In addition, real association signals can, in some cases, be removed by the technical correction. The major limitation is the inability to distinguish technical artefacts from actual association signals. In situations where the sequencing (or more generally technical) batches are fully confounded with the case control batches, no statistical methodology can separate signal from noise.

This issue is reflected by the fact that some cohorts are highly corrected with the Technical Kinship and/or the Read Depth Kinship matrices. It was thus not possible to find a single kinship or particular combination thereof that worked sufficiently well across all cohorts. By using spectral decomposition as discussed in Section 1.7, it is possible to include a single kinship that is solved in the same time as a standard linear regression. However, for multiple kinships, it is necessary to use an approximation, such as GRAMMAR. This was not perfect however, as the variance component estimations from Section 4.4.8 (page 100) show that some cohorts naively exhibit a perfect or close to perfect ($\geq 95\%$) correlation with these Kinship Matrices. The SNP filtering that was ultimately used corrected for this in some cohorts by lowering

the variance explained. In these cases the joint model retained enough power to detect the strongest signals, but some associations were presumably lost. This is to be expected however with any model that corrects for factors such as PS or technical artefacts. Over correction is less of an issue than under correction as we are interested in only the top variants i.e. an inflation factor of 0.9 is better than 1.1.

5.3 Implications for experimental design

Obviously, the ideal experimental design involves homogeneous case control cohorts from the technical standpoint. However, this is not always feasible given the costs of sequencing controls. This issue is particularly strong for rare diseases. Indeed, while for complex traits with small effect sizes, equal case control cohorts maximize the power to detect associations, the situation is different for rare diseases with large effect sizes. The optimum ratio of controls to cases is equal to the odds ratio parameter, which raises the need for large control cohorts that are ideally shared across studies. This idea has been applied very successfully by the Wellcome Trust Case Control Consortium (WTCCC) that has provided shared controls to the medical genetics community [Lee et al., 2014; Todd et al., 2007; Lindgren et al., 2009].

However, there are practical situations where the approach can be effective at correcting sub-optimal case control designs. For example, if cases were sequenced using capture technology A, and controls were sequenced using a combination of technologies A and B, the technical correction will appropriately remove variants present in excess in samples sequenced using technology A, independently of the case control status. Hence, even a limited number of controls sequenced using technology A can be sufficient to provide useful information to separate signal from background. Similarly, the addition of cases sequenced using technology B can serve the same purpose to verify that candidate variants present in technology A cases are also found in excess in technology B cases. Statistically, factors that may induce some systematic artefact, eg technology used, sample plate information, where the sample was sequenced, who prepared it etc may by partially modelled for as a random effect. However, there are limits, as the number of fixed or random effects included in the model should be limited to avoid removing all signal.

As discussed already, many possible sources of artefacts exist. PCR bias [Kanagawa, 2003], PCR

artefacts [Kurata et al., 2004] and amplification inefficiencies or drift [Walsh PS, Erlich HA, 1992; Mutter and Boynton, 1995] to name a few. In the early rounds of PCR, certain amplicons may be stochastically amplified more than others, resulting in a skewed distribution later on. As this drift bias is random, it is likely not a large factor in systematic differences between cases and controls however. Bias can also be caused by a relative difference in the size of genomes in the solution being amplified; smaller genomes have been shown to be overamplified compared to larger genomes [Pinard et al., 2006]. This last problem can be overcome by amplifying the target genome in isolation, readily achieved through methods such as microfluidic droplets and microdissections [Woyke et al., 2010].

The concentration of the DNA template is also important. It is increasingly difficult to equally amplify the entirety of a sequence as its concentration lowers [Chandler et al., 1997]. This can pose problems as research moves towards lower starting quantities of target DNA; for example in the cases of single cell or cell-free DNA [Woyke et al., 2010]. This has led to the development of Amplication free methods which lower or eliminate traditional sources of bias [Karlsson et al., 2015].

5.3.1 Remaining sources of artefacts, impact of sequence capture and transition to WGS

My analysis of exome data points to a variety of artefacts that remain difficult to control for without applying drastic quality control measures. Exploratory analysis of the data clearly shows that the artefacts identified are correlated with sequencing batches, which are in turn associated with sequence capture technologies. However, this does not imply that all these artefacts are created by the sequencing technology itself. For example, manual curation of the results of the HCM and ARVC association results (Chapter 2, Tables 3 and 4) identified several cases of C>A/G>T transversion artefacts, as described in [Costello et al., 2013], which are likely to result from oxidation of DNA during acoustic shearing in samples containing reactive contaminants from the extraction process. Two examples of this are shown in Figure 5.1. This is most likely a property of the sequencing batch, or perhaps of the sequencing facility during a specific period of time, which typically happens to be confounded with capture technology in this case. Another issue when calling

variants is the difficulty that aligners have when dealing with repetitive regions. Homopolymer sequences are often miscalled as indels, so care must be taken to ensure that this is not just due to misalignment. Figure 5.2 shows one such erroneous call. Manually examining the reads via IGV in this way is invaluable in determining if this call is real. Nevertheless, other sources of artefacts, for example capture of paralogue sequences that are mismapped to target regions, are a consequence of capture technology choices.

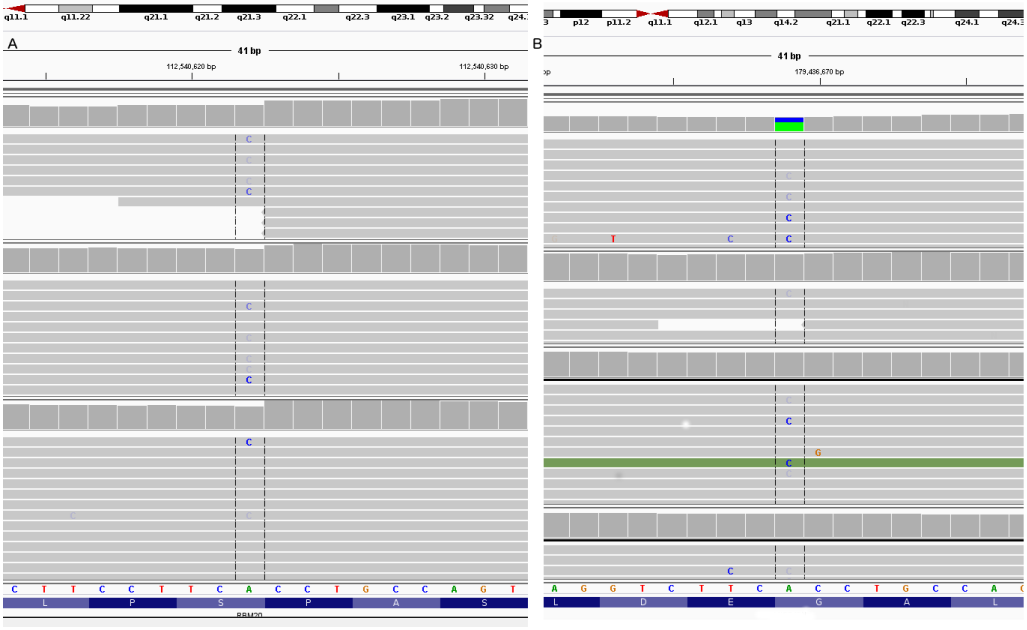


Figure 5.1: Two examples of C>A Transversion Artefacts shown in two HCM samples via IGV. On the left, a *RBM20* variant, chr10:112540622A-C, and on the right a TTN variant chr2:179436669A-C.

A conclusion from these observations is that some of these challenges will be alleviated by the transition from exome/capture sequencing to WGS, because of the removal of a potential source of differential bias between cohorts. In addition, the analysis of non-coding regions remains an unsolved challenge. Even for broad capture techniques, in the case of the HCM and ARVC targeted sequencing datasets, the higher level of polymorphisms in non coding regions combined with the absence of large reference datasets such as EXaC [Consortium et al., 2015] prevented us from obtaining meaningful association results. This is particularly frustrating as likely causal variants are detected in only half of HCM patients, and it is likely that regulatory regions contribute to at least some extent to that missing heritability.

Given the limitations highlighted above, ongoing projects such as Genomics England appear to be

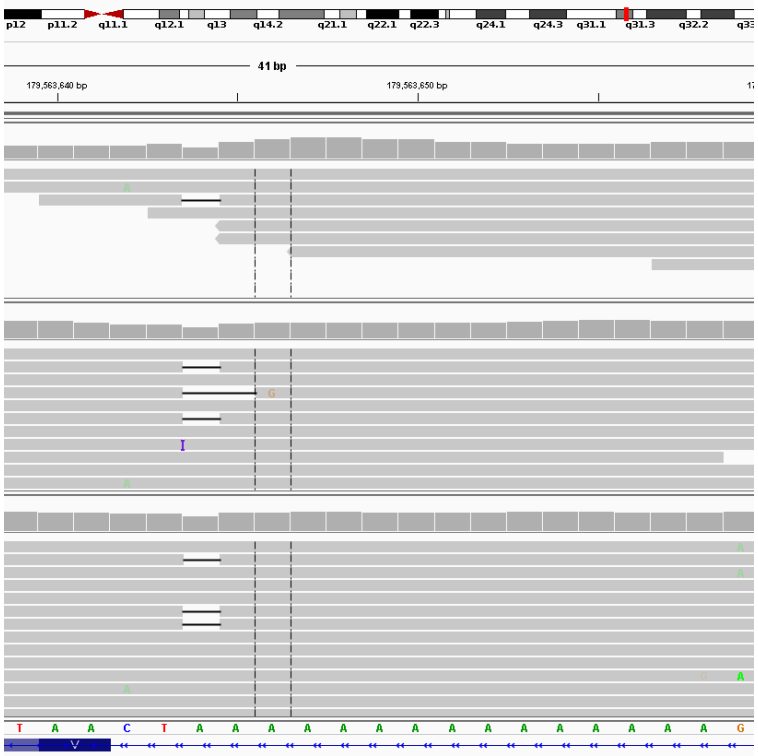


Figure 5.2: *TTN* homopolymer run. The insertion that was called here, chr2:179563646-A, is bounded by black vertical dashed lines. The DNA sequence of the region is labelled below, highlighting the extent of the homopolymer region.

a major step forward toward a refined understanding of cardiovascular disease genetics. ARVC, HCM and DCM alongside many other conditions are included in this projects remit (Figure 5.3). This project will achieve the combined aims of (i) increasing sample size to improve statistical power, (ii) reduce the technical artefacts that complicate the analysis of sequence data and cross-cohort comparison and (iii) provide a new window into the yet unexplored role of regulatory regions. These data will provide a unique opportunity to revisit and expand the outcome of exon centric association studies that are presented in this thesis.

Rare Diseases currently included in the 100,000 Genomes Project

Category	Subcategory	Disease
Cardiovascular disorders	Arteriopathies	Familial hypercholesterolaemia
	Connective Tissues Disorders and Aortopathies	Familial Thoracic Aortic Aneurysm Disease Eligibility
	Cardiac Arrhythmia	Brugada Syndrome Long QT Syndrome Catecholaminergic Polymorphic Ventricular Tachycardia
	Cardiomyopathy	Arrhythmogenic Right Ventricular Cardiomyopathy Left Ventricular Noncompaction Cardiomyopathy Dilated Cardiomyopathy (DCM) Dilated Cardiomyopathy and conduction defects Hypertrophic Cardiomyopathy
	Congenital heart disease	Fallots tetralogy Hypoplastic Left Heart Syndrome Pulmonary atresia Transposition of the great vessels Left Ventricular Outflow Tract obstruction disorders Isomerism and laterality disorders
	Lymphatic disorders	Meige disease

October 2015

The controlled copy of this document is maintained in the Genomics England internal document management system. Any copies of this document held outside of that system, in whatever format (for example, paper or email attachment), are considered to have passed out of control and should be checked for currency and validity. This document is uncontrolled when printed.

Figure 5.3: The types of cardiovascular disease that are included in the 100,000 Genomes Project. For more information on factors such as diagnostic criteria of these conditions refer to the Genomics England website.

5.4 Comparison to other methods

There are many companies that offer capture platforms for WES. The platforms that they provide have improved considerably over time. Despite this, many issues still exist, as discussed extensively already (Section 1.6.2). Some solutions have been proposed to solve these issues. Consider the case of ArtQ from the Introduction. While this worked in that specific case, it is far from a general solution. More far reaching, another approach is assigning genotypes through the use of imputation [Davies et al., 2016]. The efficacy of such an approach will vary depending on the discrete pattern of missingness, be it missing completely at random, missing at random or not missing at random. However, it has been shown to have biases too and indeed these are more pronounced for rare variants [Palmer and Pe'er, 2016]. Thus, this approach would not be ideal for the data in UCL-ex, which is dominated by rare variants.

Chapter 4 provides the first demonstration of an attempt to use alternative Kinship matrices to

control for factors other than PS. This approach is promising in that it does not require an idea of the cause of the data, unlike ArtQ. Furthermore, the concepts' derivation from standard methods of correcting PS in GWAS means it is statistically robust. However, at the time of submission of this thesis, this had not progressed sufficiently to offer adequate interpretable results. It is therefore difficult to comprehensively compare its performance against other methods. Work is ongoing, and it is being integrated into the principle pipeline for analysing WES at the UCL Genetics Institute.

An interesting alternative was proposed in [Palmer and Pe'er, 2016]. Multiple Imputation (MI) functions better than traditional imputation as it probabilistically assigns genotypes by generating posterior probabilities that are weighted by the confidence in the data. This intuitively makes sense: you would expect variants that are of lower quality to be downweighted in comparison to high quality variants. Such methods provide a benchmark against which other solutions need to be compared to identify if they provide a genuine improvement.

Appendix

.1 Chapter 2 - Cardiac Case Control

.1.1 Molecular Autopsy of Sudden Arrhythmic Death Syndrome Gene panel

	Channelopathy associated or candidate genes
1	AKAP9 (A kinase (PRKA) anchor protein (yotiao) 9)
2	ANK2 (ankyrin 2, neuronal)
3	ANKRD1 (ankyrin repeat domain 1 (cardiac muscle)
4	CACNA1B (calcium channel, voltage-dependent, N type, alpha 1B subunit)
5	CACNA1C (calcium channel, voltage-dependent, L type, alpha 1C subunit)
6	CACNA1D (calcium channel, voltage-dependent, L type, alpha 1D subunit)
7	CACNA2D1 (calcium channel, voltage-dependent, alpha 2/delta subunit 1)
8	CACNB2 (calcium channel, voltage-dependent, beta 2 subunit)
9	CASQ2 (calsequestrin 2 (cardiac muscle))
10	CAV3 (caveolin 3)
11	DPP6 (dipeptidyl-peptidase 6)
12	GJA1 (gap junction protein, alpha 1, 43kDa)
13	GJA5 (gap junction protein, alpha 5, 40kDa)
14	GPD1L (glycerol-3-phosphate dehydrogenase 1-like)
15	HCN1 (hyperpolarization activated cyclic nucleotide-gated potassium channel 1)
16	HCN4 (hyperpolarization activated cyclic nucleotide-gated potassium channel 4)
17	KCNA5 (potassium voltage-gated channel, shaker-related subfamily, member 5)
18	KCND3 (potassium voltage-gated channel, Shal-related subfamily, member 3)
19	KCNE1 (potassium voltage-gated channel, Isk-related family, member 1)
20	KCNE1L (KCNE1-like)
21	KCNE2 (potassium voltage-gated channel, Isk-related family, member 2)
22	KCNE3 (potassium voltage-gated channel, Isk-related family, member 3)
23	KCNE4 (potassium voltage-gated channel, Isk-related family, member 4)
24	KCNH2 (potassium voltage-gated channel, subfamily H (eag-related), member 2)
25	KCNJ11 (potassium inwardly-rectifying channel, subfamily J, member 11)
26	KCNJ12 (potassium inwardly-rectifying channel, subfamily J, member 12)
27	KCNJ2 (potassium inwardly-rectifying channel, subfamily J, member 2)
28	KCNJ3 (potassium inwardly-rectifying channel, subfamily J, member 3)
29	KCNJ5 (potassium inwardly-rectifying channel, subfamily J, member 5)
30	KCNJ8 (potassium inwardly-rectifying channel, subfamily J, member 8)
31	KCNQ1 (potassium voltage-gated channel, KQT-like subfamily, member 1)
32	KCNQ2 (potassium voltage-gated channel, KQT-like subfamily, member 2)
33	NPPA (natriuretic peptide A)
34	RANGRF (RAN guanine nucleotide release factor)
35	RYR2 (ryanodine receptor 2 (cardiac))
36	SCN1B (sodium channel, voltage-gated, type I, beta subunit)
37	SCN2B (sodium channel, voltage-gated, type II, beta subunit)
38	SCN3B (sodium channel, voltage-gated, type III, beta subunit)
39	SCN4B (sodium channel, voltage-gated, type IV, beta subunit)
40	SCN5A (sodium channel, voltage-gated, type V, alpha subunit)
41	SCNN1B (sodium channel, non-voltage-gated 1, beta subunit)
42	SCNN1G (sodium channel, non-voltage-gated 1, gamma subunit)
43	SNTA1 (syntrophin, alpha 1)
00	Cardiomyopathy_associated_or_candidate_genes
01	ABCC9 (ATP-binding cassette, sub-family C member 9)
02	ACTC1 (Actin, alpha cardiac muscle 1)
03	ACTN2 (Actinin, alpha 2)
03	AGL (amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase)
04	BAG3 (BCL2-associated athanogene 3)
05	BRAF (v-raf murine sarcoma viral oncogene homolog B1)

Table 1: Name according to HGNC of the candidate genes for the Molecular Autopsy of Sudden Cardiac Death study

	Cardiomyopathy associated or candidate genes
1	CALR3 (calreticulin 3)
2	CRYAB (crystallin, alpha B)
3	CSRP3 (cysteine and glycine-rich protein 3 (cardiac LIM protein)
4	DES (desmin)
5	DMD (dystrophin)
6	DSC2 (desmocolin 2)
7	DSG2 (desmoglein 2)
8	DSP (desmoplakin)
9	DTNA (dystrobrevin, alpha)
10	EMD (emerin)
11	EYA4 (eyes absent homolog 4)
12	FHL1 (four and a half LIM domains 1)
13	FHL2 (four and a half LIM domains 2)
14	FKTN (fukutin)
15	FLNC (filamin C, gamma)
16	FXN (frataxin)
17	GAA (glucosidase, alpha; acid)
18	GATAD1 (GATA zinc finger domain containing 1)
19	GLA (galactosidase, alpha)
20	HRAS (v-Ha-ras Harvey rat sarcoma viral oncogene homolog)
21	ILK (integrin-linked kinase)
22	JPH2 (junctophilin 2)
23	JUP (junction plakoglobin)
24	KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog)
25	LAMA4 (laminin, alpha 4)
26	LAMP2 (lysosomal-associated membrane protein 2)
27	LDB3 (LIM domain binding 3)
28	LMNA (lamin A/C)
29	MAP2K1 (mitogen-activated protein kinase kinase 1)
30	MYBPC3 (myosin binding protein C, cardiac)
31	MYH6 (myosin, heavy polypeptide 6, cardiac muscle, alpha)
32	MYH7 (myosin, heavy polypeptide 7, cardiac muscle, alpha)
33	MYL2 (myosin, light chain 2, regulatory, cardiac, slow)
34	MYL3 (myosin, light chain 3, alkali; ventricular, skeletal, slow)
35	MYLK2 (myosin light chain kinase 2)
36	MYOT (myotilin)
37	MYOZ2 (myozenin 2)
38	MYPN (myopalladin)
39	NEBL (nebulette)
40	NEXN (nexilin (F actin binding protein))
41	NRAS (neuroblastoma RAS viral (v-ras) oncogene homolog)
42	PDLIM3 (PDZ and LIM domain 3)
43	PKP2 (plakophilin 2)
44	PLEC (plectin)
45	PKP4 (plakophilin 4)
46	PLN (phospholamban)
47	PNN (pinin, desmosome associated protein)
48	PRKAG2 (protein kinase, AMP-activated, gamma 2 non-catalytic subunit)
49	PSEN1 (presenilin 1)
50	PSEN2 (presenilin 2)

Table 2: Name according to HGNC of the candidate genes for the Molecular Autopsy of Sudden Cardiac Death study

	Cardiomyopathy associated or candidate genes
1	PTPN11 (protein tyrosine phosphatase, non-receptor type 11)
2	RAF1 (v-raf-1 murine leukemia viral oncogene homolog 1)
3	RBM20 (RNA binding motif protein 20)
4	SGCD (sarcoglycan, delta (35kDa dystrophin-associated glycoprotein))
5	SHOC2 (soc-2 suppressor of clear homolog (C. elegans))
6	SLC25A4 (solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4)
7	SOS1 (son of sevenless homolog 1 (Drosophila))
8	TAZ (tafazzin)
9	TCAP (titin-cap)
10	TGFB3 (transforming growth factor, beta 3)
11	TMEM43 (transmembrane protein 43)
12	TMPO (thymopoietin)
13	TNNC1 (troponin C type 1 (slow))
14	TNNI3 (troponin I type 3 (cardiac))
15	TNNT2 (troponin T type 2 (cardiac))
16	TPM1 (tropomyosin 1 (alpha))
17	TTN (titin)
18	TTR (transthyretin)
19	VCL (vinculin)
00	Others
01	ADRB2 (adrenoceptor beta 1)
02	ADRB2 (adrenoceptor beta 2)
03	ADRB3 (adrenoceptor beta 3)
04	BMPR2 (bone morphogenetic protein receptor, type II (serine/threonine kinase)
05	CTF1 (cardiotrophin 1)
06	DNM1L (dynamin 1-like)
07	ELN (elastin)
08	GATA4 (GATA binding protein 4)
09	potassium inwardly-rectifying channel, subfamily J, member 11
10	LRP6 (low density lipoprotein receptor-related protein 6)
11	NKX2-5 (NK2 homeobox 5)
12	TBX20 (T-box 20)
13	FBN1 (fibrillin 1)
14	FBN2 (fibrillin 2)
15	TGFBR1 (transforming growth factor, beta receptor I)
16	TGFBR2 (transforming growth factor, beta receptor II)
17	ACTA2 (actin, alpha 2, smooth muscle, aorta)

Table 3: Name according to HGNC of the candidate genes for the Molecular Autopsy of Sudden Cardiac Death study

.1.2 ARVC/HCM Gene Panel

	ENSEMBL	HUGO	Chromosome	Start	End
1	ENSG00000118194	TNNT2	1	201359008	201377762
2	ENSG00000118729	CASQ2	1	115700007	115768781
3	ENSG00000159166	LAD1	1	201373244	201399915
4	ENSG00000160789	LMNA	1	156082573	156140089
5	ENSG00000198626	RYR2	1	237042205	237833988
6	ENSG00000144283	PKP4	2	158456964	158682879
7	ENSG00000153237	CCDC148	2	158171081	158456753
8	ENSG00000155657	TTN	2	178525989	178830802
9	ENSG00000175084	DES	2	219418377	219426739
10	ENSG00000237298	TTN-AS1	2	178521183	178779963
11	ENSG00000114854	TNNC1	3	52451102	52454070
12	ENSG00000125046	SSUH2	3	8619400	8745040
13	ENSG00000160808	MYL3	3	46857872	46882169
14	ENSG00000170876	TMEM43	3	14124940	14143679
15	ENSG00000182533	CAV3	3	8733800	8841808
16	ENSG00000183873	SCN5A	3	38548057	38649673
17	ENSG00000145362	ANK2	4	112818109	113383740
18	ENSG00000154553	PDLIM3	4	185500660	185535612
19	ENSG00000096696	DSP	6	7541575	7586717
20	ENSG00000152661	GJA1	6	121435692	121449727
21	ENSG00000198523	PLN	6	118548298	118560730
22	ENSG00000055118	KCNH2	7	150944961	150978315
23	ENSG00000178209	PLEC	8	143915147	143976734
24	ENSG00000035403	VCL	10	73995193	74121363
25	ENSG00000122367	LDB3	10	86668449	86736068
26	ENSG00000203867	RBM20	10	110644397	110839469
27	ENSG00000053918	KCNQ1	11	2444684	2849109
28	ENSG00000129170	CSRP3	11	19182030	19210573
29	ENSG00000134571	MYBPC3	11	47331397	47352702
30	ENSG00000057294	PKP2	12	32790745	32896840
31	ENSG00000111245	MYL2	12	110910819	110920722
32	ENSG00000092054	MYH7	14	23412738	23435718
33	ENSG00000100941	PNN	14	39175183	39183218
34	ENSG00000119699	TGFB3	14	75958099	75982991
35	ENSG00000197616	MYH6	14	23381990	23408277
36	ENSG00000259083	RP11-407N17.4	14	39174885	39175880
37	ENSG00000140416	TPM1	15	63042632	63071915
38	ENSG00000159251	ACTC1	15	34788096	34796139
39	ENSG00000123700	KCNJ2	17	70168673	70180048
40	ENSG00000173801	JUP	17	41754604	41786931
41	ENSG00000173991	TCAP	17	39664187	39666555
42	ENSG00000046604	DSG2	18	31498043	31549008
43	ENSG00000134755	DSC2	18	31058840	31102415
44	ENSG00000129991	TNNI3	19	55151767	55157773
45	ENSG00000267110	CTD-2587H24.4	19	55154757	55160671
46	ENSG00000159197	KCNE2	21	34364024	34371389
47	ENSG00000180509	KCNE1	21	34446688	34512275

Table 4: Genes sequenced in the ARVC/HCM Gene panel

.2 Chapter 3 - HCM Copy Number Variant analysis gene panel

Gene	Ensembl Number	Chromosome:Start-End	Number(bp)
MYBPC3	ENSG00000134571	Chr11:47352958-47374253	21295
MYH7	ENSG00000092054	Chr14:23881948-23904870	22922
TNNI3	ENSG00000129991	Chr19:55663137-55669100	5963
TNNT2	ENSG00000118194	Chr1:201328143-201346805	18662
TPM1	ENSG00000140416	Chr15:63334838-63364111	29273
MYL2	ENSG00000111245	Chr12:111348626-111358404	9778
MYL3	ENSG00000160808	Chr3:46899357-46904973	5616
ACTC1	ENSG00000159251	Chr15:35080297-35087927	7630
TNNC1	ENSG00000114854	Chr3:52485108-52488057	2949
MYH6	ENSG00000197616	Chr14:23851199-23877482	26283
TTN	ENSG00000155657	Chr2:179390720-179672150	281430
PDLIM3	ENSG00000154553	Chr4:186422852-186456712	33860
CSRP3	ENSG00000129170	Chr11:19203578-19223589	20011
DES	ENSG00000175084	Chr2:220283099-220291459	8360
LMNA	ENSG00000160789	Chr1:156084461-156109878	25417
LDB3	ENSG00000122367	Chr10:88428426-88495822	67396
VCL	ENSG00000035403	Chr10:75757872-75879912	122040
TCAP	ENST00000309889	Chr17:37821599-37822806	1207
PLN	ENSG00000198523	Chr6:118869442-118881586	12144
RBM20	ENSG00000203867	Chr10:112404155-112599227	195072
JUP	ENSG00000173801	Chr17:39910859-39942964	32105
DSP	ENSG00000096696	Chr6:7541870-7586946	45076
PKP2	ENSG00000057294	Chr12:32943682-33049780	106098
DSG2	ENSG00000046604	Chr18:29078027-29128813	50786
DSC2	ENSG00000134755	Chr18:28645944-28682388	36444
RYR2	ENSG00000198626	Chr1:237205702-237997288	791586
TMEM43	ENST00000306077	Chr3:14166440-14185180	18740
TGF-3	ENST00000238682	Chr14:76424442-76448092	23650
KCNQ1	ENSG00000053918	Chr11:2466221-2870339	404118
KCNH2	ENSG00000055118	Chr7:150642050-150675014	32964
SCN5A	ENSG00000183873	Chr3:38589554-38691164	101610
KCNE1	ENSG00000180509	Chr21:35818989-35828063	9074
KCNE2	ENSG00000159197	Chr21:35736323-35743440	7117
ANK2	ENST00000394537	Chr4:113970785-114304894	334109
CASQ2	ENSG00000118729	Chr1:116242628-116311426	68798
CAV3	ENSG00000182533	Chr3:8775496-8788450	12954
KCNJ2	ENSG00000123700	Chr17:68165676-68176181	10505
PLEC	ENSG00000178209	Chr8:144989321-145025044	35723
GJA1	ENST00000282561	Chr6:121756745-121770872	14127
PKP4	ENSG00000144283	Chr2:159313476-159537938	224462
PNN	ENSG00000100941	Chr14:39644387-39652421	8036

Table 5: Name of the targeted genes, Ensembl accesion number, chromsomal positon and size sequenced for the HCM CNV study.

.3 Chapter 4 - Single Variant Results for additional UCLex Cohorts

For each of the tables in this section the layout is the same so, a common legend is provided: SNP details the position of the tested variant (hg19). Gene is the HUGO name for the gene in which the SNP resides. FisherPvalue is the pvalue from Fisher's exact test. LRpvalue is the Linear Regression pvalue with no covariates or kinship matrices. LMMpvalue is the pvalue from Equation 4.9. OR is the risk odds ratio. 'Homs' are homozygotes for the minor allele, while 'Hets' are heteroozygotes.

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Huntingtons	nb.Homs.ctrls	nb.Hets.Huntingtons	nb.Hets.ctrls
1	1_7723921_T_C	CAMTA1	8.82E-04	1.00E-16	1.00E-16	81	0	0	2	2
2	18_72223597_-_TGT	CNDP1	1.19E-12	1.00E-16	1.00E-16	NA	1	0	4	0
3	3_49845468_C_T	UBA7	8.46E-04	1.00E-16	1.00E-16	83	0	0	2	2
4	11_44228488_C_T	EXT2	9.16E-04	1.00E-16	1.87E-11	80	0	0	2	2

Table 6: Huntingtons Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Eye	nb.Homs.ctrls	nb.Hets.Eye	nb.Hets.ctrls
1	10_100189567_C_G	HPS1	8.70E-12	1.00E-16	1.00E-16	358	0	0	7	1
2	10_115368200_C_T	NRAP	9.47E-13	1.00E-16	1.00E-16	NA	0	0	7	0
3	10_123970354_A_G	TACC2	2.97E-09	1.00E-16	1.00E-16	NA	0	0	5	0
4	10_124753444_A_G	IKZF5	3.30E-09	1.00E-16	1.00E-16	NA	0	0	5	0
5	10_16992007_G_C	CUBN	2.23E-06	1.00E-16	1.00E-16	100	0	0	4	2
6	10_27524067_T_C	ACBD5	8.82E-07	1.00E-16	1.00E-16	192	0	0	4	1
7	10_3193452_G_A	PITRM1	3.70E-05	1.00E-16	1.00E-16	139	0	0	3	1
8	10_84745067_A_G	NRG3	9.58E-13	1.00E-16	1.00E-16	NA	0	0	7	0
9	11_100211919_T_C	CNTN5	2.12E-07	1.00E-16	1.00E-16	NA	0	0	4	0
10	11_433357_G_-	ANO9	7.45E-06	1.00E-16	1.00E-16	25	0	0	5	9
11	1_150530506_G_T	ADAMTSL4	7.74E-09	1.00E-16	1.00E-16	296	0	0	5	1
12	1_153279608_C_T	PGLYRP3	3.62E-09	1.00E-16	1.00E-16	NA	0	0	5	0
13	1_154920148_G_A	PBXIP1	8.66E-10	1.00E-16	1.00E-16	70	0	0	7	5
14	11_56237921_C_T	OR5M3	6.56E-11	1.00E-16	1.00E-16	NA	0	0	6	0
15	11_56237921_C_T	OR8U8	6.56E-11	1.00E-16	1.00E-16	NA	0	0	6	0
16	1_158670285_G_C	OR6K2	3.32E-11	1.00E-16	1.00E-16	183	0	0	7	2
17	1_171605478_G_A	MYOC	5.80E-07	1.00E-16	1.00E-16	52	0	0	5	5
18	1_174210750_G_A	RABGAP1L	1.64E-04	1.00E-16	1.00E-16	48	0	0	3	3
19	1_182353781_A_C	GLUL	6.62E-07	1.00E-16	1.00E-16	206	0	0	4	1
20	11_85445453_T_C	SYTL2	3.27E-05	1.00E-16	1.00E-16	145	0	0	3	1

Table 7: Opthalmology Condition 1 Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	NA..3	nb.Homs.ctrls	NA..4	nb.Hets.ctrls
1	11_120175740_C_T	POU2F3	8.51E-09	1.00E-16	1.00E-16	NA	0	0	4	0
2	22_29621158_C_T	EMID1	6.14E-11	1.00E-16	1.00E-16	NA	0	0	5	0
3	22_30688584_C_T	TBC1D10A	6.64E-11	1.00E-16	1.00E-16	NA	0	0	5	0
4	2_56420484_C_A	CCDC85A	5.38E-06	1.00E-16	1.00E-16	272	0	0	3	1
5	3_169700534_A_C	SEC62	9.23E-10	1.00E-16	1.00E-16	466	0	0	5	1
6	8_25234859_G_A	DOCK5	9.78E-09	1.00E-16	1.00E-16	NA	0	0	4	0
7	11_2432745_G_A	TRPM5	9.77E-04	1.00E-16	2.45E-16	67	0	0	2	3
8	16_51175457_C_T	SALL1	3.32E-06	1.00E-16	5.19E-16	320	0	0	3	1
9	17_17124804_C_T	FLCN	8.16E-06	1.00E-16	1.08E-15	161	0	0	3	2
10	3_38307622_G_A	SLC22A13	8.37E-06	1.00E-16	3.42E-15	159	0	0	3	2
11	19_334440_T_G	MIER2	7.43E-06	1.00E-16	1.47E-13	166	0	0	3	2
12	1_53932322_A_G	DMRTB1	1.58E-04	1.00E-16	9.16E-13	36	0	0	3	9
13	1_55014014_G_A	ACOT11	1.23E-05	1.00E-16	2.06E-12	140	0	0	3	2
14	12_55420621_G_A	NEUROD4	7.54E-06	1.00E-16	9.22E-12	165	0	0	3	2
15	9_19785979_G_A	SLC24A2	1.16E-07	1.00E-16	1.18E-11	215	0	0	4	2
16	1_889212_G_A	NOC2L	1.55E-07	1.00E-16	2.38E-11	200	0	0	4	2
17	1_114394645_A_C	PTPN22	7.54E-06	1.00E-16	3.67E-11	165	0	0	3	2
18	1_201868510_G_T	LMOD1	1.10E-05	1.00E-16	3.97E-11	146	0	0	3	2
19	17_9631505_C_T	USP43	6.72E-04	3.90E-15	4.27E-11	94	0	0	2	2
20	1_979517_T_A	AGRN	2.74E-07	1.00E-16	4.32E-11	142	0	0	4	3

Table 8: Icelandic IBD Cohort Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Neur	nb.Homs.ctrls	nb.Hets.Neur	nb.Hets.ctrls
1	1_109806800_A_G	CELSR2	6.12E-07	1.00E-16	1.00E-16	NA	0	0	4	0
2	11_19077074_CT_GT	MRGPRX2	1.13E-09	1.00E-16	1.00E-16	254	0	0	6	1
3	11_19077075_T_G	MRGPRX2	3.18E-10	1.00E-16	1.00E-16	NA	0	0	6	0
4	1_161695685_TACG_GACG	FCRLB	1.26E-10	1.00E-16	1.00E-16	240	0	0	7	1
5	11_6977031_A_G	ZNF215	1.12E-06	1.00E-16	1.00E-16	NA	0	0	4	0
6	11_700213_G_A	TMEM80	1.13E-06	1.00E-16	1.00E-16	NA	0	0	4	0
7	11_76928315_-_ATCT	GDPD4	2.61E-06	1.00E-16	1.00E-16	145	0	0	4	1
8	12_123342763_G_A	HIP1R	5.42E-07	1.00E-16	1.00E-16	NA	0	0	4	0
9	13_25671955_T_C	PABPC3	1.80E-06	1.00E-16	1.00E-16	NA	0	0	4	0
10	13_39425226_G_T	FREM2	3.27E-08	1.00E-16	1.00E-16	NA	0	0	5	0
11	14_33291745_G_A	AKAP6	5.36E-07	1.00E-16	1.00E-16	NA	0	0	4	0
12	14_88893017_C_T	SPATA7	6.47E-07	1.00E-16	1.00E-16	NA	0	0	4	0
13	16_84270704_C_T	KCNG4	1.15E-06	1.00E-16	1.00E-16	NA	0	0	4	0
14	17_2227024_C_G	SRR	1.59E-08	1.00E-16	1.00E-16	NA	0	0	4	0
15	17_2227024_C_G	TSR1	1.59E-08	1.00E-16	1.00E-16	NA	0	0	4	0
16	17_39346627_CCACCCAACA_-	KRTAP9-1	2.78E-19	1.00E-16	1.00E-16	156	0	0	14	3
17	17_39346639_CTGTCAAACC_-	KRTAP9-1	3.84E-19	1.00E-16	1.00E-16	152	0	0	14	3
18	1_75037184_C_G	C1orf173	2.71E-08	1.00E-16	1.00E-16	NA	0	0	5	0
19	17_72366771_T_G	GPR142	1.01E-06	1.00E-16	1.00E-16	NA	0	0	4	0
20	19_20002909_A_C	ZNF253	5.68E-07	1.00E-16	1.00E-16	79	0	0	5	2

Table 9: Neurology Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	NA..3	nb.Homs.ctrls	NA..4	nb.Hets.ctrls
1	9_140878697_C_T	CACNA1B	6.99E-04	6.83E-15	1.21E-16	70	0	1	2	3
2	15_42192873_C_T	EHD4	2.01E-04	1.00E-16	1.81E-16	174	0	0	2	2
3	1_175365772_G_A	TNR	6.93E-06	1.25E-15	6.83E-13	123	1	0	1	5
4	19_48714997_G_A	CARD8	3.77E-04	2.32E-06	3.24E-10	101	1	0	0	4

Table 10: Opthalmology Condition 2 Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Skin	nb.Homs.ctrls	nb.Hets.Skin	nb.Hets.ctrls
1	10_121140398_A_G	GRK5	5.47E-07	1.00E-16	1.00E-16	143	0	0	4	2
2	10_128923854_A_G	DOCK1	1.62E-09	1.00E-16	1.00E-16	414	0	0	5	1
3	10_129216722_A_G	DOCK1	5.22E-13	1.00E-16	1.00E-16	343	0	0	7	2
4	10_135099022_G_T	TUBGCP2	1.79E-08	1.00E-16	1.00E-16	NA	0	0	4	0
5	10_16975189_C_T	CUBN	3.31E-05	1.00E-16	1.00E-16	99	0	0	3	2
6	10_31799735_A_G	ZEB1	6.24E-08	1.00E-16	1.00E-16	NA	0	0	4	0
7	10_38406867_G_T	ZNF37A	2.54E-06	1.00E-16	1.00E-16	72	0	0	4	4
8	10_72181468_C_T	EIF4EBP2	8.99E-04	5.73E-15	1.00E-16	80	0	0	2	2
9	10_73562724_G_A	CDH23	1.43E-16	1.00E-16	1.00E-16	420	0	0	9	2
10	10_79553803_C_T	DLG5	2.17E-11	1.00E-16	1.00E-16	24	0	0	11	38
11	10_79576826_C_T	DLG5	5.12E-12	1.00E-16	1.00E-16	28	1	0	9	29
12	10_79616631_C_T	DLG5	4.35E-12	1.00E-16	1.00E-16	28	1	0	9	34
13	10_82403828_TGT_-	SH2D4B	3.45E-15	1.00E-16	1.00E-16	NA	0	0	8	0
14	10_90537864_G_C	LIPN	4.04E-10	1.00E-16	1.00E-16	67	1	0	5	8
15	10_90575223_T_C	LIPM	1.54E-10	1.00E-16	1.00E-16	86	1	0	5	6
16	10_93904826_A_G	CPEB3	1.85E-11	1.00E-16	1.00E-16	159	1	0	5	3
17	10_95275274_A_G	CEP55	7.43E-09	1.00E-16	1.00E-16	41	1	0	5	10
18	1_109823574_A_G	PSRC1	1.80E-05	1.00E-16	1.00E-16	122	0	0	3	2
19	11_107375857_G_T	ALKBH8	2.33E-05	1.00E-16	1.00E-16	112	0	0	3	2
20	11_114401546_G_A	NXPE1	4.98E-11	1.00E-16	1.00E-16	119	0	0	7	4

Table 11: Dermatology Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	NA..3	nb.Homs.ctrls	NA..4	nb.Hets.ctrls
1	10_106209864_G_T	CCDC147	4.52E-04	1.00E-16	1.00E-16	81	0	0	2	9
2	10_117704227_A_G	ATRNL1	3.67E-04	1.00E-16	1.00E-16	91	0	1	2	6
3	10_1421303_A_C	ADARB2	5.51E-06	1.00E-16	1.00E-16	121	1	0	1	9
4	10_1421304_G_C	ADARB2	5.51E-06	1.00E-16	1.00E-16	121	1	0	1	9
5	10_24833905_C_T	KIAA1217	7.82E-04	1.00E-16	1.00E-16	59	1	0	0	12
6	10_29581461_A_G	LYZL1	3.75E-04	1.00E-16	1.00E-16	90	0	0	2	8
7	10_3823777_T_C	KLF6	7.12E-05	1.00E-16	1.00E-16	299	0	0	2	2
8	10_3824081_G_A	KLF6	5.34E-05	1.00E-16	1.00E-16	347	0	0	2	2
9	10_43596103_G_A	RET	1.41E-04	1.00E-16	1.00E-16	169	0	1	2	2
10	10_75184902_G_A	MSS51	1.38E-06	1.00E-16	1.00E-16	220	1	0	1	5
11	10_84744970_C_T	NRG3	3.46E-06	1.00E-16	1.00E-16	145	1	0	1	8
12	10_91477375_G_T	KIF20B	4.68E-04	1.00E-16	1.00E-16	81	0	0	2	8
13	10_97154762_G_A	SORBS1	5.26E-04	1.00E-16	1.00E-16	74	0	0	2	10
14	1_100387183_T_A	AGL	1.46E-04	1.00E-16	1.00E-16	166	0	0	2	4
15	1_10709186_G_A	CASZ1	1.31E-04	1.00E-16	1.00E-16	176	1	0	0	4
16	1_109803697_G_A	CELSR2	8.29E-06	1.00E-16	1.00E-16	103	0	0	3	11
17	11_100141950_G_A	CNTN5	2.54E-04	1.00E-16	1.00E-16	112	0	0	2	10
18	11_102196019_A_G	BIRC3	8.31E-05	1.00E-16	1.00E-16	240	0	0	2	3
19	11_107375667_C_T	ALKBH8	1.47E-04	1.00E-16	1.00E-16	166	0	0	2	4
20	11_111753245_C_T	C11orf1	6.34E-04	1.00E-16	1.00E-16	66	0	1	2	9

Table 12: Keratoconus Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Immune	nb.Homs.ctrls	nb.Hets.Immune	nb.Hets.ctrls
1	10_48429512_G_A	GDF10	2.26E-07	1.00E-16	1.00E-16	148	0	0	5	1
2	1_197072458_T_C	ASPM	1.62E-06	1.00E-16	1.00E-16	NA	0	0	4	0
3	12_68720470_A_G	MDM1	3.07E-04	4.29E-13	1.00E-16	45	0	0	3	2
4	19_11565623_G_T	ELAVL3	1.20E-06	1.00E-16	1.00E-16	NA	1	0	2	0
5	1_9787030_G_A	PIK3CD	7.85E-10	1.00E-16	1.00E-16	NA	0	0	6	0
6	20_25258960_T_C	PYGB	6.65E-04	2.53E-10	1.00E-16	29	0	0	3	3
7	2_127453624_G_A	GYPC	5.42E-06	1.00E-16	1.00E-16	119	1	0	2	1
8	2_241463616_G_A	ANKMY1	3.29E-04	3.05E-13	1.00E-16	44	0	0	3	2
9	2_46588218_C_T	EPAS1	6.41E-04	1.35E-10	1.00E-16	29	0	0	3	3
10	3_51671458_G_A	RAD54L2	2.91E-04	5.39E-14	1.00E-16	46	0	0	3	2
11	6_33283594_G_A	ZBTB22	5.59E-05	1.04E-11	1.00E-16	16	0	0	5	9
12	6_33372831_T_C	KIFC1	2.17E-05	1.37E-11	1.00E-16	13	0	0	6	13
13	6_38906754_T_C	DNAH8	3.50E-04	8.22E-13	1.00E-16	43	0	0	3	2
14	6_38906754_T_C	LOC100131047	3.50E-04	8.22E-13	1.00E-16	43	0	0	3	2
15	7_150918769_G_A	ABCF2	3.07E-04	2.79E-13	1.00E-16	45	0	0	3	2
16	X_10085293_G_A	WWC3	2.95E-04	1.00E-16	1.00E-16	46	0	0	3	2
17	16_57095409_G_A	NLRC5	2.05E-08	1.00E-16	1.03E-16	43	1	0	5	5
18	14_103450076_G_A	CDC42BPB	6.84E-06	1.00E-16	1.10E-16	112	0	0	4	1
19	3_48658942_C_T	TMEM89	3.13E-07	1.00E-16	3.26E-16	36	1	0	4	6
20	1_12433865_C_T	VPS13D	4.51E-04	6.95E-09	3.95E-16	14	0	0	4	8

Table 13: Primary Immuno Deficiency Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	NA..3	nb.Homs.ctrls	NA..4	nb.Hets.ctrls
1	1_152681693_TGTGGT_-	LCE4A	7.93E-75	1.00E-16	1.00E-16	189	3	0	52	10
2	X_73811755_C_G	RLIM	1.85E-60	5.52E-16	5.91E-13	59	0	0	57	27
3	7_150325310_C_T	GIMAP6	9.24E-04	4.27E-03	2.64E-12	31	0	0	3	2
4	1_248616401_G_A	OR2T2	1.95E-38	1.34E-05	9.55E-11	6	2	21	101	279
5	1_248616408_C_T	OR2T2	5.41E-37	6.23E-05	1.21E-10	5	2	22	102	299
6	2_10584626_C_T	ODC1	3.83E-04	2.05E-01	1.53E-10	6	1	0	5	21

Table 14: Prion Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Mito	nb.Homs.ctrls	nb.Hets.Mito	nb.Hets.ctrls
1	10_120934107_-_A	PRDX3	5.84E-08	1.00E-16	1.00E-16	127	0	0	5	2
2	1_153314126_C_T	PGLYRP4	1.75E-04	1.00E-16	1.00E-16	47	0	0	3	3
3	12_111772320_C_T	CUX2	3.57E-05	1.00E-16	1.00E-16	141	0	0	3	1
4	12_15650326_T_C	PTPRO	5.12E-05	1.00E-16	1.00E-16	125	0	0	3	1
5	1_228595950_C_T	TRIM17	1.73E-04	1.00E-16	1.00E-16	47	0	0	3	3
6	12_55759555_C_T	OR6C75	3.50E-05	1.00E-16	1.00E-16	142	0	0	3	1
7	14_21841524_A_G	SUPT16H	1.02E-04	1.00E-16	1.00E-16	67	0	0	3	2
8	17_65925556_A_G	BPTF	8.78E-05	1.00E-16	1.00E-16	70	0	0	3	2
9	19_24102851_A_G	ZNF726	2.08E-04	1.00E-16	1.00E-16	44	0	0	3	3
10	1_982833_C_T	AGRN	2.86E-09	1.00E-16	1.00E-16	NA	0	0	5	0
11	2_108994856_C_T	SULT1C4	4.23E-05	1.00E-16	1.00E-16	133	0	0	3	1
12	2_170092467_G_A	LRP2	3.57E-04	7.75E-15	1.00E-16	33	0	0	3	4
13	3_138187558_C_T	ESYT3	1.49E-04	1.00E-16	1.00E-16	49	0	0	3	3
14	3_193081064_G_A	ATP13A5	3.68E-05	1.00E-16	1.00E-16	139	0	0	3	1
15	3_58855204_C_T	C3orf67	1.95E-07	1.00E-16	1.00E-16	80	0	0	5	3
16	4_13616292_T_A	BOD1L1	1.46E-04	1.00E-16	1.00E-16	50	0	0	3	3
17	5_72980694_C_T	RGNEF	6.05E-05	1.00E-16	1.00E-16	80	0	0	3	2
18	5_75989260_C_T	IQGAP2	7.58E-05	1.00E-16	1.00E-16	74	0	0	3	2
19	5_90136800_A_C	GPR98	3.19E-05	1.00E-16	1.00E-16	147	0	0	3	1
20	7_100635250_C_A	MUC12	3.52E-04	4.51E-13	1.00E-16	31	0	0	3	5

Table 15: Mitochondrial disease Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	nb.Homs.Bone	nb.Homs.ctrls	nb.Hets.Bone	nb.Hets.ctrls
1	11_66099992_G_T	RIN1	3.81E-07	1.00E-16	1.00E-16	NA	0	0	5	0
2	1_170955836_C_T	C1orf129	9.79E-06	1.00E-16	1.00E-16	NA	0	0	4	0
3	16_3614342_G_A	NLRC3	7.16E-04	2.59E-10	1.00E-16	49	0	0	3	1
4	5_146763533_A_G	STK32A	7.12E-04	1.50E-10	1.00E-16	49	0	0	3	1
5	9_117139815_G_A	AKNA	9.83E-06	1.59E-10	1.00E-16	18	1	0	4	7
6	11_66334727_C_T	CTSF	7.28E-06	1.16E-15	1.52E-16	45	0	0	5	2
7	3_72495945_A_G	RYBP	7.24E-04	4.88E-10	3.40E-16	49	0	0	3	1
8	22_50187923_G_T	BRD1	4.59E-04	1.67E-10	1.63E-15	58	0	0	3	1
9	9_123673632_C_T	TRAF1	9.77E-06	1.00E-16	1.07E-14	NA	0	0	4	0
10	5_156381625_C_T	TIMD4	7.44E-04	1.60E-10	1.67E-14	48	0	0	3	1
11	7_4830898_C_G	AP5Z1	6.77E-04	1.42E-10	2.87E-14	50	0	0	3	1
12	7_44180307_G_A	MYL7	5.83E-04	9.08E-11	7.23E-14	53	0	0	3	1
13	11_65146965_A_G	SLC25A45	6.59E-04	2.12E-10	1.13E-13	51	0	0	3	1
14	10_102739999_C_T	SEMA4G	6.49E-04	1.54E-10	1.39E-13	51	0	0	3	1
15	10_102739999_C_T	MRPL43	6.49E-04	1.54E-10	1.39E-13	51	0	0	3	1
16	15_65983590_A_G	DENND4A	2.81E-04	1.94E-08	2.94E-13	12	0	0	5	7
17	5_141694394_C_T	SPRY4	1.50E-04	1.08E-09	1.07E-12	86	0	0	3	1
18	X_100169508_G_A	XKRX	5.13E-05	2.64E-14	1.12E-12	65	0	0	4	1
19	7_35293222_T_A	TBX20	1.64E-04	1.10E-11	1.29E-12	32	0	0	4	2
20	19_16006368_G_A	CYP4F2	7.09E-06	1.00E-16	1.32E-12	NA	0	0	4	0

Table 16: Bone Marrow Failure Single Variant Results

	SNP	Gene	FisherPvalue	LRpvalue	LMMpvalue	OR	NA..3	nb.Homs.ctrls	NA..4	nb.Hets.ctrls
1	19_36002488_C_A	DMKN	6.23E-04	1.00E-16	1.00E-16	97	0	0	2	2
2	19_36027710_C_T	GAPDHS	8.59E-04	1.00E-16	2.29E-16	71	0	0	2	3
3	2_235951605_C_T	SH3BP4	8.35E-06	1.00E-16	7.68E-15	160	1	0	1	2
4	17_39115095_G_A	KRT39	3.08E-07	1.00E-16	9.71E-14	437	1	0	1	3
5	16_71483003_C_T	ZNF23	3.21E-06	1.00E-16	1.45E-13	323	0	0	3	1
6	1_11188142_C_T	MTOR	8.05E-04	2.95E-15	2.51E-12	74	1	0	0	3
7	1_152192053_C_T	HRNR	7.93E-04	1.00E-16	7.07E-12	74	0	0	2	3
8	5_132652228_G_A	FSTL4	7.93E-04	1.00E-16	1.64E-11	74	0	0	2	3
9	1_17396685_G_A	PADI2	7.82E-04	1.00E-16	3.07E-11	75	0	0	2	3
10	5_156923974_C_T	ADAM19	1.01E-05	1.00E-16	5.70E-11	39	0	0	4	11
11	19_55086356_-_GT	LILRA2	5.17E-05	1.00E-16	2.08E-10	59	0	0	3	6
12	19_55086359_GC_-	LILRA2	7.40E-05	1.00E-16	4.02E-10	50	0	0	3	7
13	4_111539617_T_A	PITX2	7.77E-04	1.00E-16	9.52E-10	75	0	0	2	3

Table 17: Opthalmology Condition 3 Single Variant Results

Bibliography

Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard a Gibbs, Matt E Hurles, and Gil a McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, oct 2010. ISSN 1476-4687. doi: 10.1038/nature09534. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3042601{&}tool=pmcentrez{&}rendertype=abstract`.

Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark a DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil a McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, nov 2012. ISSN 1476-4687. doi: 10.1038/nature11632. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3498066{&}tool=pmcentrez{&}rendertype=abstract`.

Daniel Aird, Wei-Shen Chen, Michael Ross, Kristen Connolly, Jim Meldrim, Carsten Russ, Sheila Fisher, David Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing bias in Illumina sequencing libraries. *Genome Biology*, 11(Suppl 1):P3, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-s1-p3.

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009. doi: 10.1101/gr.094052.109.vidual.

L Almasy and J Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *American journal of human genetics*, 62(5):1198–1211, 1998. ISSN 00029297. doi: 10.1086/301844.

C I Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American journal of human genetics*, 54(3):535–543, 1994. ISSN 0002-9297.

Ivan Angulo, Oscar Vadas, Fabien Garçon, Edward Banham-Hall, Vincent Plagnol, Timothy R Leahy, Helen Baxendale, Tanya Coulter, James Curtis, Changxin Wu, Katherine Blake-Palmer, Olga Perisic, Deborah Smyth, Mailis Maes, Christine Fiddler, Jatinder Juss, Deirdre Cilliers, Gašper Markelj, Anita Chandra, George Farmer, Anna Kielkowska, Jonathan Clark, Sven Kracker, Marianne Debré, Capucine Picard, Isabelle Pellier, Nada Jabado, James a Morris, Gabriela Barcenas-Morales, Alain Fischer, Len Stephens, Phillip Hawkins, Jeffrey C Barrett, Mario Abinun, Menna Clatworthy, Anne Durandy, Rainer Doffinger, Edwin R Chilvers, Andrew J Cant, Dinakantha Kumararatne, Klaus Okkenhaug, Roger L Williams, Alison Condliffe, and Sergey Nejentsev. Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage. *Science (New York, N.Y.)*, 342(6160):866–71, nov 2013. ISSN 1095-9203. doi: 10.1126/science.1243292. URL `http://www.ncbi.nlm.nih.gov/pubmed/24136356`.

David Araújo-Vilar, Joaquin Lado-Abeal, Fernando Palos-Paz, Giovanna Lattanzi, Manuel A Bandín, Diego Bellido, Lourdes Domínguez-Gerpe, Carlos Calvo, Oscar Pérez, Alia Ramazanova, Noelia Martínez-Sánchez, Berta Victoria, and Ana Teresa Costa-Freitas. A novel phenotypic expression associated with a new mutation in LMNA gene, characterized by partial lipodystrophy, insulin resistance, aortic stenosis and hypertrophic cardiomyopathy. *Clinical endocrinology*, 69(1):61–8, jul 2008. ISSN 1365-2265. doi: 10.1111/j.1365-2265.2007.03146.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/18031308`.

William Astle and David J. Balding. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4):451–471, nov 2009. ISSN 0883-4237. doi: 10.1214/09-STS307. URL `http://projecteuclid.org/euclid.ss/1271770342`.

Paul L Auer, Guillaume Lettre, AR Wood, T Esko, J Yang, S Vedantam, TH Pers, S Gustafsson, JA Tennessen, AW Bigham, TD O, JR Huyghe, AU Jackson, MP Fogarty, ML Buchkovich, A Stancakova, HM Stringham, JC Cohen, RS Kiss, A Pertsemlidis, YL Marcel, R McPherson, HH Hobbs, J Cohen, A Pertsemlidis, IK Kotowski, R Graham, CK Garcia, HH Hobbs, JC Cohen, E Boerwinkle, TH Mosley,

HH Hobbs, EM Roth, JM McKenney, C Hanotin, G Asset, EA Stein, EA Stein, S Mellis, GD Yancopoulos, N Stahl, D Logan, WB Smith, MP Dolled-Filhart, M Lee, CW Ou-Yang, RR Haraksingh, JC Lin, Y Li, C Sidore, HM Kang, M Boehnke, GR Abecasis, C Gilissen, A Hoischen, HG Brunner, JA Veltman, W Fu, TD O, M Beaudoin, KS Lo, A N, MA Rivas, M Beaudoin, A Gardet, C Stevens, Y Sharma, CK Zhang, D Diogo, F Kurreeman, EA Stahl, KP Liao, N Gupta, JD Greenberg, X Zhan, DE Larson, C Wang, DC Koboldt, YV Sergeev, RS Fulton, J Kozlitina, E Smagris, S Stender, BG Nordestgaard, HH Zhou, A Tybjaerg-Hansen, GM Peloso, PL Auer, JC Bis, A Voorman, AC Morrison, NO Stitziel, OL Holmen, H Zhang, Y Fan, DH Hovelson, EM Schmidt, W Zhou, PL Auer, A Teumer, U Schick, A O, TA Manolio, FS Collins, NJ Cox, DB Goldstein, LA Hindorff, DJ Hunter, V Agarwala, J Flannick, S Sunyaev, TDC Go, D Altshuler, GV Kryukov, A Shpunt, JA Stamatoyannopoulos, SR Sunyaev, D Li, JP Lewinger, WJ Gauderman, CE Murcray, D Conti, LT Guey, J Kravic, O Melander, NP Burtt, JM Laramie, V Lyssenko, LA Lange, Y Hu, H Zhang, C Xue, EM Schmidt, ZZ Tang, MJ Emond, T Louie, J Emerson, W Zhao, RA Mathias, MR Knowles, J Flannick, G Thorleifsson, NL Beer, SB Jacobs, N Grarup, NP Burtt, IJ Barnett, S Lee, X Lin, DY Lin, D Zeng, ZZ Tang, A Helgason, S Sigurethardottir, J Nicholson, B Sykes, EW Hill, DG Bradley, K Hatzikotoulas, A Gilly, E Zeggini, J Gudmundsson, P Sulem, DF Gudbjartsson, G Masson, BA Agnarsson, KR Benediktsdottir, V Steinthorsdottir, G Thorleifsson, P Sulem, H Helgason, N Grarup, A Sigurdsson, E Levy-Lahad, R Catane, S Eisenberg, B Kaufman, G Hornreich, E Lishinsky, SB Ng, AW Bigham, KJ Buckingham, MC Hannibal, MJ McMillin, HI Gildersleeve, SB Ng, KJ Buckingham, C Lee, AW Bigham, HK Tabor, KM Dent, ET Cirulli, DB Goldstein, H Hu, JC Roach, H Coon, SL Guthery, KV Voelkerding, RL Margraf, O Zuk, SF Schaffner, K Samocha, R Do, E Hechter, S Kathiresan, RS Spielman, RE McGinnis, WJ Ewens, PC Sham, SM Purcell, NM Laird, C Lange, D Altshuler, LD Brooks, A Chakravarti, FS Collins, MJ Daly, P Donnelly, V Bansal, O Libiger, A Torkamani, NJ Schork, S Morgenthaler, WG Thilly, B Li, SM Leal, AL Price, GV Kryukov, PI Bakker, SM Purcell, J Staples, LJ Wei, MC Wu, S Lee, T Cai, Y Li, M Boehnke, X Lin, S Lee, MJ Emond, MJ Bamshad, KC Barnes, MJ Rieder, DA Nickerson, S Lee, GR Abecasis, M Boehnke, X Lin, I Mathieson, G McVean, TD O, MC Babron, M Tayrac, DN Rutledge, E Zeggini, E Genin, Q Liu, DL Nicolae,

LS Chen, CJ Willer, Y Li, GR Abecasis, DJ Liu, GM Peloso, X Zhan, OL Holmen, M Zawistowski, S Feng, S Lee, TM Teslovich, M Boehnke, X Lin, J Marchini, B Howie, SR Browning, BL Browning, B Howie, C Fuchsberger, M Stephens, J Marchini, GR Abecasis, Y Li, CJ Willer, J Ding, P Scheet, GR Abecasis, J Marchini, B Howie, S Myers, G McVean, P Donnelly, SI Berndt, S Gustafsson, R Magi, A Ganna, E Wheeler, MF Feitosa, TM Teslovich, K Musunuru, AV Smith, AC Edmondson, IM Stylianou, M Koseki, PL Auer, JM Johnsen, AD Johnson, BA Logsdon, LA Lange, MA Nalls, M Du, PL Auer, S Jiao, J Haessler, D Altshuler, E Boerwinkle, V Orru, M Steri, G Sole, C Sidore, F Virdis, M Dei, SI Vrieze, SM Malone, U Vaidyanathan, A Kwong, HM Kang, X Zhan, Q Duan, EY Liu, PL Auer, G Zhang, EM Lange, G Jun, G Pistis, E Porcu, SI Vrieze, C Sidore, M Steri, F Danjou, DG MacArthur, S Balasubramanian, A Frankish, N Huang, J Morris, K Walter, ET Lim, P Wurtz, AS Havulinna, P Palta, T Tukiainen, K Rehnstrom, GV Kryukov, LA Pennacchio, SR Sunyaev, M Kircher, DM Witten, P Jain, BJ O, MT Maurano, R Humbert, E Rynes, RE Thurman, E Haugen, H Wang, R Andersson, C Gebhard, I Miguel-Escalada, I Hoof, J Bornholdt, M Boyd, EV Davydov, DL Goode, M Sirota, GM Cooper, A Sidow, S Batzoglou, LD Ward, M Kellis, A Javed, S Agrawal, PC Ng, IA Adzhubei, S Schmidt, L Peshkin, VE Ramensky, A Gerasimova, P Bork, AP Boyle, EL Hong, M Hariharan, Y Cheng, MA Schaub, M Kasowski, S Petrovski, Q Wang, EL Heinzen, AS Allen, DB Goldstein, P Kumar, S Henikoff, PC Ng, W McLaren, B Pritchard, D Rios, Y Chen, P Flicek, F Cunningham, R Calabrese, E Capriotti, P Fariselli, PL Martelli, R Casadio, X Liu, X Jian, E Boerwinkle, AR Majithia, J Flannick, P Shahinian, M Guo, MA Bray, P Fontanillas, KJ Karczewski, JT Dudley, KR Kukurba, R Chen, AJ Butte, SB Montgomery, MA Schaub, AP Boyle, A Kundaje, S Batzoglou, M Snyder, G Trynka, C Sandor, B Han, H Xu, BE Stranger, XS Liu, KS Lo, S Vadlamudi, MP Fogarty, KL Mohlke, G Lettre, GS Barsh, GP Copenhaver, G Gibson, SM Williams, S Sanna, B Li, A Mulas, C Sidore, HM Kang, AU Jackson, A Kamb, S Harper, K Stefansson, S Purcell, SS Cherny, and PC Sham. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16, 2015. ISSN 1756-994X. doi: 10.1186/s13073-015-0138-2. URL `http://genomemedicine.com/content/7/1/16`.

Yurii S. Aulchenko, Dirk Jan De Koning, and Chris Haley. Genomewide rapid association using mixed

model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, 2007. ISSN 00166731. doi: 10.1534/genetics.107.075614.

Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa

Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Charles Lee, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C. L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Stephen T. Sherry, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Bur-

chard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Gonçalo R. Abecasis, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Richard M. Durbin, Matthew E. Hurles, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Senduran Balasubramaniam, Yuan Chen, Vincenza Colonna, Petr Danecek, Luke Jostins, Thomas M. Keane, Shane McCarthy, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Yingrui Li, Ruibang Luo, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Steven A. McCarroll, Robert E. Handsaker, David M. Altshuler, Eric Banks, Guillermo del Angel, Giulio Genovese, Chris Hartl, Heng Li, Seva Kashin, James C. Nemesh, Khalid Shakir, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Jeremiah Degenhardt, Jan O. Korbel, Markus H. Fritz, Sascha Meiers, Benjamin Raeder, Tobias Rausch, Adrian M. Stütz, Paul Flicek, Francesco Paolo Casale, Laura Clarke, Richard E. Smith, Oliver Stegle, Xiangqun Zheng-Bradley, David R. Bentley, Bret Barnes, R. Keira Cheetham, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, Richard Shaw, Eric-Wubbo Lameijer, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Li Ding, Ira Hall, Kai Ye, Phil Lacroute, Charles Lee, Eliza Cerveira, Ankit Malhotra, Jaeho Hwang, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, David W. Craig, Nils Homer, Deanna Church, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Vineet Bafna, Jacob Michaelson, Kenny

Ye, Scott E. Devine, Eugene J. Gardner, Gonçalo R. Abecasis, Jeffrey M. Kidd, Ryan E. Mills, Gargi Dayama, Sarah Emery, Goo Jun, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Mark J. Chaisson, Fereydoun Hormozdiari, John Huddleston, Maika Malig, Bradley J. Nelson, Peter H. Sudmant, Nicholas F. Parrish, Ekta Khurana, Matthew E. Hurles, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Alexej Abyzov, Jieming Chen, Declan Clarke, Hugo Lam, Xinmeng Jasmine Mu, Cristina Sisu, Jing Zhang, Yan Zhang, Richard A. Gibbs, Fuli Yu, Matthew Bainbridge, Danny Challis, Uday S. Evani, Christie Kovar, James Lu, Donna Muzny, Uma Nagaswamy, Jeffrey G. Reid, Aniko Sabo, Jin Yu, Xiaosen Guo, Wangshen Li, Yingrui Li, Renhua Wu, Gabor T. Marth, Erik P. Garrison, Wen Fung Leong, Alistair N. Ward, Guillermo del Angel, Mark A. DePristo, Stacey B. Gabriel, Namrata Gupta, Chris Hartl, Ryan E. Poplin, Andrew G. Clark, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Richard E. Smith, Xiangqun Zheng-Bradley, Daniel G. MacArthur, Elaine R. Mardis, Robert Fulton, Daniel C. Koboldt, Simon Gravel, Carlos D. Bustamante, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Stephen T. Sherry, Chunlin Xiao, Emmanouil T. Dermitzakis, Gonçalo R. Abecasis, Hyun Min Kang, Gil A. McVean, Mark B. Gerstein, Suganthi Balasubramanian, Lukas Habegger, Haiyuan Yu, Paul Flicek, Laura Clarke, Fiona Cunningham, Ian Dunham, Daniel Zerbino, Xiangqun Zheng-Bradley, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Stephen B. Montgomery, Marianne K. DeGorter, Ekta Khurana, Chris Tyler-Smith, Yuan Chen, Vincenza Colonna, Yali Xue, Mark B. Gerstein, Suganthi Balasubramanian, Yao Fu, Donghoon Kim, Adam Auton, Anthony Marcketta, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Erik P. Garrison, Robert E. Handsaker, Seva Kashin, Steven A. McCarroll, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Carlos D. Bustamante, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, Charles Lee, Eliza Cerveira, Ankit Malhotra, Mallory Romanovitch, Chengsheng Zhang, Gonçalo R. Abecasis, Lachlan Coin, Haojing Shao, David Mittelman, Chris Tyler-Smith, Qasim Ayub, Ruby Banerjee, Maria Cerezo, Yuan Chen, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Shane McCarthy, Graham R. Ritchie, Yali Xue,

Fengtang Yang, Richard A. Gibbs, Christie Kovar, Divya Kalra, Walker Hale, Donna Muzny, Jeffrey G. Reid, Jun Wang, Xu Dan, Xiaosen Guo, Guoqing Li, Yingrui Li, Chen Ye, Xiaole Zheng, David M. Altshuler, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, David R. Bentley, Anthony Cox, Sean Humphray, Scott Kahn, Ralf Sudbrak, Marcus W. Albrecht, Matthias Lienhard, David Larson, David W. Craig, Tyler Izatt, Ahmet A. Kurdoglu, Stephen T. Sherry, Chunlin Xiao, David Haussler, Gonçalo R. Abecasis, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, Thomas M. Keane, Shane McCarthy, James Stalker, Aravinda Chakravarti, Bartha M. Knoppers, Gonçalo R. Abecasis, Kathleen C. Barnes, Christine Beiswanger, Esteban G. Burchard, Carlos D. Bustamante, Hongyu Cai, Hongzhi Cao, Richard M. Durbin, Norman P. Gerry, Neda Gharani, Richard A. Gibbs, Christopher R. Gignoux, Simon Gravel, Brenna Henn, Danielle Jones, Lynn Jorde, Jane S. Kaye, Alon Keinan, Alastair Kent, Angeliki Kerasidou, Yingrui Li, Rasika Mathias, Gil A. McVean, Andres Moreno-Estrada, Pilar N. Ossorio, Michael Parker, Alissa M. Resch, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Ralf Sudbrak, Zhongming Tian, Sarah Tishkoff, Lorraine H. Toji, Chris Tyler-Smith, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Andres Ruiz-Linares, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Taras K. Oleksyk, Kathleen C. Barnes, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Pardis C. Sabeti, Jiayong Zhu, Xiaoyan Deng, Pardis C. Sabeti, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Trâ‘n Tnh Hiê‘n, Sarah J. Dunstan, Nguyen Thuy Hang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, Pardis C. Sabeti, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Eric D. Green, Audrey Duncanson, Michael Dunn, Jeffery A. Schloss, Jun Wang, Huanming Yang, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, sep 2015. ISSN 0028-0836.

doi: 10.1038/nature15393. URL http://www.nature.com/doifinder/10.1038/nature15393.

Richard D Bagnall, Laura Yeates, and Christopher Semsarian. The role of large gene deletions and duplications in MYBPC3 and TNNT2 in patients with hypertrophic cardiomyopathy. *International journal of cardiology*, 145(1):150–3, nov 2010. ISSN 1874-1754. doi: 10.1016/j.ijcard.2009.07.009. URL http://www.ncbi.nlm.nih.gov/pubmed/19666196.

Jingru Bao, Jizheng Wang, Yan Yao, Yilu Wang, Xiaohan Fan, Kai Sun, Ding Sheng He, Frank I Marcus, Shu Zhang, Rutai Hui, and Lei Song. Correlation of ventricular arrhythmias with genotype in arrhythmogenic right ventricular cardiomyopathy. *Circulation. Cardiovascular genetics*, 6(6):552–6, dec 2013. ISSN 1942-3268. doi: 10.1161/CIRCGENETICS.113.000122. URL http://www.ncbi.nlm.nih.gov/pubmed/24125834.

E R Behr, a Casey, M Sheppard, M Wright, T J Bowker, M J Davies, W J McKenna, and D a Wood. Sudden arrhythmic death syndrome: a national survey of sudden unexplained cardiac death. *Heart (British Cardiac Society)*, 93(5):601–605, 2007. ISSN 1355-6037. doi: 10.1136/hrt.2006.099598.

Erez Ben-Yaacov and Yonina C Eldar. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics (Oxford, England)*, 24(16):i139–45, aug 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn272. URL http://www.ncbi.nlm.nih.gov/pubmed/18689815.

Katharine M Brauch, Margaret L Karst, Kathleen J Herron, Mariza de Andrade, Patricia A Pellikka, Richard J Rodeheffer, Virginia V Michels, and Timothy M Olson. Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy. *Journal of the American College of Cardiology*, 54(10):930–41, sep 2009. ISSN 1558-3597. doi: 10.1016/j.jacc.2009.05.038. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2782634{&}tool=pmcentrez{&}rendertype=abstract.

Francesca Brun, Carl V Barnes, Gianfranco Sinagra, Dobromir Slavov, Giulia Barbati, Xiao Zhu, Sharon L Graw, Anita Spezzacatene, Bruno Pinamonti, Marco Merlo, Ernesto E Salcedo, William H Sauer, Matthew R G Taylor, and Luisa Mestroni. Titin and desmosomal genes in the natural history of arrhythmogenic

right ventricular cardiomyopathy. *Journal of medical genetics*, 51(10):669–76, oct 2014. ISSN 1468-6244. doi: 10.1136/jmedgenet-2014-102591. URL http://jmg.bmj.com/content/51/10/669.long{#}ref-10.

Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. *Lancet*, 361 (9357):598–604, feb 2003. ISSN 0140-6736. doi: 10.1016/S0140-6736(03)12520-2. URL http://www.ncbi.nlm.nih.gov/pubmed/12598158.

Elisa Carniel, Matthew R G Taylor, Gianfranco Sinagra, Andrea Di Lenarda, Lisa Ku, Pamela R Fain, Mark M Boucek, Jean Cavanaugh, Snjezana Miocic, Dobromir Slavov, Sharon L Graw, Jennie Feiger, Xiao Zhong Zhu, Dmi Dao, Debra a Ferguson, Michael R Bristow, and Luisa Mestroni. Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. *Circulation*, 112(1):54–9, jul 2005. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.104.507699. URL http://www.ncbi.nlm.nih.gov/pubmed/15998695.

Marina Cerrone and Mario Delmar. Desmosomes and the sodium channel complex: Implications for arrhythmogenic cardiomyopathy and Brugada syndrome. *Trends in cardiovascular medicine*, pages 1–7, mar 2014. ISSN 1873-2615. doi: 10.1016/j.tcm.2014.02.001. URL http://www.ncbi.nlm.nih.gov/pubmed/24656989.

V Chanavat, M F Seronde, P Bouvagnet, P Chevalier, R Rousson, and G Millat. Molecular characterization of a large MYBPC3 rearrangement in a cohort of 100 unrelated patients with hypertrophic cardiomyopathy. *European journal of medical genetics*, 55(3):163–6, mar 2012. ISSN 1878-0849. doi: 10.1016/j.ejmg.2012.01.002. URL http://www.ncbi.nlm.nih.gov/pubmed/22314326.

D P Chandler, J K Fredrickson, and F J Brockman. Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular ecology*, 6(5):475–82, may 1997. ISSN 0962-1083. URL http://www.ncbi.nlm.nih.gov/pubmed/9161015.

Navin Chandra, Rachel Bastiaenen, Michael Papadakis, and Sanjay Sharma. Sudden cardiac death in young athletes: practical challenges and diagnostic dilemmas. *Journal of the American College of Cardiology*, 61 (10):1027–40, mar 2013. ISSN 1558-3597. doi: 10.1016/j.jacc.2012.08.1032.

Chandra Sekhar Reddy Chilamakuri, Susanne Lorenz, Mohammed-Amin Madoui, Daniel Vodák, Jinchang Sun, Eivind Hovig, Ola Myklebost, and Leonardo a Meza-Zepeda. Performance comparison of four exome capture systems for deep sequencing. *BMC genomics*, 15:449, jan 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-449. URL http://www.ncbi.nlm.nih.gov/pubmed/24912484.

V. Chirico, V. Ferraù, I. Loddo, S. Briuglia, M. Amorini, V. Salpietro, A. Lacquaniti, C. Salpietro, and T. Arrigo. LMNA gene mutation as a model of cardiometabolic dysfunction: From genetic analysis to treatment response. *Diabetes & Metabolism*, 40(3):224–228, jun 2014. ISSN 12623636. doi: 10.1016/j.diabet.2013.12.008. URL http://linkinghub.elsevier.com/retrieve/pii/S1262363614000020.

Yong Won Choi, Tae Jun Park, Hyo Soo Kim, and In Kyoung Lim. Signals regulating necrosis of cardiomyoblast by BTG2(/TIS21/PC3) via activation of GSK3β and opening of mitochondrial permeability transition pore in response to H2O2. *Biochemical and biophysical research communications*, 434(3):559–65, may 2013. ISSN 1090-2104. doi: 10.1016/j.bbrc.2013.03.114. URL http://www.ncbi.nlm.nih.gov/pubmed/23583382.

Jennifer D Churchill, Sara J Bowne, Lori S Sullivan, Richard Alan Lewis, Dianna K Wheaton, David G Birch, Kari E Branham, John R Heckenlively, and Stephen P Daiger. Mutations in the X-linked retinitis pigmentosa genes RPGR and RP2 found in 8.5% of families with a provisional diagnosis of autosomal dominant retinitis pigmentosa. *Investigative ophthalmology & visual science*, 54(2):1411–6, feb 2013. ISSN 1552-5783. doi: 10.1167/iovs.12-11541. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3597192{&}tool=pmcentrez{&}rendertype=abstract.

David G Clayton, Neil M Walker, Deborah J Smyth, Rebecca Pask, Jason D Cooper, Lisa M Maier, Luc J Smink, Alex C Lam, Nigel R Ovington, Helen E Stevens, Sarah Nutland, Joanna M M Howson, Malek Faham, Martin Moorhead, Hywel B Jones, Matthew Falkowski, Paul Hardenbol, Thomas D Willis, and John a Todd. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics*, 37(11):1243–1246, 2005. ISSN 1061-4036. doi: 10.1038/ng1653.

Donald F. Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts,

T. Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G. MacArthur, Jeffrey R. MacDonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P. Carter, Charles Lee, Stephen W. Scherer, and Matthew E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010. ISSN 0028-0836. doi: 10.1038/nature08516. URL `http://www.nature.com/doifinder/10.1038/nature08516`.

Exome Aggregation Consortium, Monkol Lek, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell, Anne O'Donnell-Luria, James Ware, Andrew Hill, Beryl Cummings, Taru Tukiainen, Daniel Birnbaum, Jack Kosmicki, Laramie Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, David Cooper, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina Peloso, Ryan Poplin, Manuel Rivas, Valentin Ruano-Rubio, Douglas Ruderfer, Khalid Shakir, Peter Stenson, Christine Stevens, Brett Thomas, Grace Tiao, Maria Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Elosua Roberto, Jose Florez, Stacey Gabriel, Gad Getz, Christina Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark McCarthy, Dermot McGovern, Ruth McPherson, Benjamin Neale, Aarno Palotie, Shaun Purcell, Danish Saleheen, Jeremiah Scharf, Pamela Sklar, Sullivan Patrick, Jaakko Tuomilehto, Hugh Watkins, James Wilson, Mark Daly, and Daniel MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. Technical report, oct 2015. URL `http://biorxiv.org/content/early/2015/10/30/030338.abstract`.

Domenico Corrado and Gaetano Thiene. Arrhythmogenic right ventricular cardiomyopathy/dysplasia: clinical impact of molecular genetic studies. *Circulation*, 113(13):1634–7, apr 2006. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.105.616490. URL `http://www.ncbi.nlm.nih.gov/pubmed/16585401`.

Maura Costello, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, Jennifer L. Fostel, Dennis C. Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B. Gabriel,

Eric S. Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):1–12, 2013. ISSN 03051048. doi: 10.1093/nar/gks1443.

Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F Conrad, Eleni Giannoulatou, Chris Holmes, Jonathan L Marchini, Kathy Stirrups, Martin D Tobin, Louise V Wain, Chris Yau, Jan Aerts, Tariq Ahmad, T Daniel Andrews, Hazel Arbury, Anthony Attwood, Adam Auton, Stephen G Ball, Anthony J Balmforth, Jeffrey C Barrett, Inês Barroso, Anne Barton, Amanda J Bennett, Sanjeev Bhaskar, Katarzyna Blaszczyk, John Bowes, Oliver J Brand, Peter S Braund, Francesca Bredin, Gerome Breen, Morris J Brown, Ian N Bruce, Jaswinder Bull, Oliver S Burren, John Burton, Jake Byrnes, Sian Caesar, Chris M Clee, Alison J Coffey, John M C Connell, Jason D Cooper, Anna F Dominiczak, Kate Downes, Hazel E Drummond, Darshna Dudakia, Andrew Dunham, Bernadette Ebbs, Diana Eccles, Sarah Edkins, Cathryn Edwards, Anna Elliot, Paul Emery, David M Evans, Gareth Evans, Steve Eyre, Anne Farmer, I Nicol Ferrier, Lars Feuk, Tomas Fitzgerald, Edward Flynn, Alistair Forbes, Liz Forty, Jayne A Franklyn, Rachel M Freathy, Polly Gibbs, Paul Gilbert, Omer Gokumen, Katherine Gordon-Smith, Emma Gray, Elaine Green, Chris J Groves, Detelina Grozeva, Rhian Gwilliam, Anita Hall, Naomi Hammond, Matt Hardy, Pile Harrison, Neelam Hassanali, Husam Hebaishi, Sarah Hines, Anne Hinks, Graham A Hitman, Lynne Hocking, Eleanor Howard, Philip Howard, Joanna M M Howson, Debbie Hughes, Sarah Hunt, John D Isaacs, Mahim Jain, Derek P Jewell, Toby Johnson, Jennifer D Jolley, Ian R Jones, Lisa A Jones, George Kirov, Cordelia F Langford, Hana Lango-Allen, G Mark Lathrop, James Lee, Kate L Lee, Charlie Lees, Kevin Lewis, Cecilia M Lindgren, Meeta Maisuria-Armer, Julian Maller, John Mansfield, Paul Martin, Dunecan C O Massey, Wendy L McArdle, Peter McGuffin, Kirsten E McLay, Alex Mentzer, Michael L Mimmack, Ann E Morgan, Andrew P Morris, Craig Mowat, Simon Myers, William Newman, Elaine R Nimmo, Michael C O'Donovan, Abiodun Onipinla, Ifejinelo Onyiah, Nigel R Ovington, Michael J Owen, Kimmo Palin, Kirstie Parnell, David Pernet, John R B Perry, Anne Phillips, Dalila Pinto, Natalie J Prescott, Inga Prokopenko, Michael A Quail, Suzanne Rafelt, Nigel W Rayner, Richard Redon, David M Reid, Renwick, Susan M Ring, Neil

Robertson, Ellie Russell, David St Clair, Jennifer G Sambrook, Jeremy D Sanderson, Helen Schuilenburg, Carol E Scott, Richard Scott, Sheila Seal, Sue Shaw-Hawkins, Beverley M Shields, Matthew J Simmonds, Debbie J Smyth, Elilan Somaskantharajah, Katarina Spanova, Sophia Steer, Jonathan Stephens, Helen E Stevens, Millicent A Stone, Zhan Su, Deborah P M Symmons, John R Thompson, Wendy Thomson, Mary E Travers, Clare Turnbull, Armand Valsesia, Mark Walker, Neil M Walker, Chris Wallace, Margaret Warren-Perry, Nicholas A Watkins, John Webster, Michael N Weedon, Anthony G Wilson, Matthew Woodburn, B Paul Wordsworth, Allan H Young, Eleftheria Zeggini, Nigel P Carter, Timothy M Frayling, Charles Lee, Gil McVean, Patricia B Munroe, Aarno Palotie, Stephen J Sawcer, Stephen W Scherer, David P Strachan, Chris Tyler-Smith, Matthew A Brown, Paul R Burton, Mark J Caulfield, Alastair Compston, Martin Farrall, Stephen C L Gough, Alistair S Hall, Andrew T Hattersley, Adrian V S Hill, Christopher G Mathew, Marcus Pembrey, Jack Satsangi, Michael R Stratton, Jane Worthington, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem Ouwehand, Miles Parkes, Nazneen Rahman, John A Todd, Nilesh J Samani, and Peter Donnelly. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289): 713–20, 2010. ISSN 1476-4687. doi: 10.1038/nature08979. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2892339{&}tool=pmcentrez{&}rendertype=abstract`.

Darshan Dalal, Khurram Nasir, Chandra Bomma, Kalpana Prakasa, Harikrishna Tandri, Jonathan Piccini, Ariel Roguin, Crystal Tichnell, Cynthia James, Stuart D Russell, Daniel P Judge, Theodore Abraham, Philip J Spevak, David a Bluemke, and Hugh Calkins. Arrhythmogenic right ventricular dysplasia: a United States experience. *Circulation*, 112(25):3823–32, dec 2005. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.105.542266. URL `http://www.ncbi.nlm.nih.gov/pubmed/16344387`.

Robert W Davies, Jonathan Flint, Simon Myers, and Richard Mott. Rapid genotype imputation from sequence without reference panels. *Nature genetics*, 48(8):965–969, aug 2016. ISSN 1546-1718. doi: 10.1038/ng.3594. URL `http://www.ncbi.nlm.nih.gov/pubmed/27376236`.

Rafael De Cid, Rabah Ben Yaou, Carinne Roudaut, Karine Charton, Sylvain Baulande, France Leturcq,

Norma Beatriz Romero, Edoardo Malfatti, Maud Beuvin, Anna Vihola, Audrey Criqui, Isabelle Nelson, Juliette Nectoux, Laurène Ben Aim, Christophe Caloustian, Robert Olaso, Bjarne Udd, Gisèle Bonne, Bruno Eymard, and Isabelle Richard. A new titinopathy: Childhood-juvenile onset Emery-Dreifuss-like phenotype without cardiomyopathy. *Neurology*, 85(24):2126–35, dec 2015. ISSN 1526-632X. doi: 10.1212/WNL.0000000000002200. URL http://www.ncbi.nlm.nih.gov/pubmed/26581302.

Dirk J de Jong, Barbara Franke, Anton H J Naber, Judith J H T Willemen, Angelien J G A M Heister, Han G Brunner, Carolien G F de Kovel, and Frans A Hol. No evidence for involvement of IL-4R and CD11B from the IBD1 region and STAT6 in the IBD2 region in Crohn's disease. *European journal of human genetics : EJHG*, 11(11):884–7, nov 2003. ISSN 1018-4813. doi: 10.1038/sj.ejhg.5201058. URL http://www.ncbi.nlm.nih.gov/pubmed/14571275.

a P W de Roos, B J Hayes, and M E Goddard. Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–53, dec 2009. ISSN 1943-2631. doi: 10.1534/genetics.109.104935. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2787438{&}tool=pmcentrez{&}rendertype=abstract.

Mark a DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011. ISSN 1061-4036. doi: 10.1038/ng.806.

B Devlin and Kathryn Roeder. Genomic Control for Association. *Biometrics*, 55(4):997–1004, 1999.

Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16):e105, sep 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn425. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2532726{&}tool=pmcentrez{&}rendertype=abstract.

Junbo Duan, Ji-Gang Zhang, Hong-Wen Deng, and Yu-Ping Wang. Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies. *PLoS ONE*, 8(3):e59128, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0059128. URL `http://dx.plos.org/10.1371/journal.pone.0059128`.

Georgios K Efthimiadis, Efstathios D Pagourelias, Thomas Gossios, and Thomas Zegkos. Hypertrophic cardiomyopathy in 2013: Current speculations and future perspectives. *World journal of cardiology*, 6(2): 26–37, feb 2014. ISSN 1949-8462. doi: 10.4330/wjc.v6.i2.26.

B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-Calling of Automated Sequencer Traces Using Phred.I. AccuracyAssessment. *Genome Research*, 8(3):175–185, mar 1998. ISSN 1088-9051. doi: 10.1101/gr.8.3.175. URL `http://genome.cshlp.org/cgi/doi/10.1101/gr.8.3.175`.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–87, aug 2003. ISSN 0016-6731. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462648{&}tool=pmcentrez{&}rendertype=abstract`.

Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature reviews. Genetics*, 7(2):85–97, feb 2006. ISSN 1471-0056. doi: 10.1038/nrg1767.

Tasha E. Fingerlin, Michael Boehnke, Gonçalo R. Abecasis, GR Abecasis, SS Cherny, WO Cookson, LR Cardon, GR Abecasis, WO Cookson, LR Cardon, GR Abecasis, E Noguchi, A Heinzmann, JA Traherne, S Bhattacharyya, NI Leaves, GG Anderson, Y Zhang, NJ Lench, A Carey, LR Cardon, MF Moffatt, WO Cookson, N Arnheim, C Strange, H Erlich, LF Barcellos, W Klitz, LL Field, R Tobias, AM Bowcock, R Wilson, MP Nelson, J Nagatomi, G Thomson, M Boehnke, CD Langefeld, NJ Camp, A Gutin, V Abkevich, JM Farnham, L Cannon-Albright, A Thomas, LR Cardon, JI Bell, CS Carlson, MA Eberle, MJ Rieder, JD Smith, L Kruglyak, DA Nickerson, CC Davis, WM Brown, EM Lange, SS Rich, CD Langefeld, E Dawson, GR Abecasis, S Bumpstead, Y Chen, S Hunt, DM Beare, J Pabial, Et Al., SB Gabriel, SF Schaffner, H Nguyen, JM Moore, J Roy, B Blumenstiel, J Higgins, M DeFelice, A Lochner, M Fag-

gart, SN Liu-Cordero, C Rotimi, A Adeyemo, R Cooper, R Ward, ES Lander, MJ Daly, D Altshuler, RC Go, MC King, J Bailey-Wilson, RC Elston, HT Lynch, AM Goldstein, RW Haile, ML Marazita, A Paganini-Hill, JBS Haldane, JM Hall, L Friedman, C Guenther, MK Lee, JL Weber, DM Black, MC King, Y Horikawa, N Oda, NJ Cox, X Li, M Orho-Melander, M Hara, Y Hinokio, Et Al., UK Kim, E Jorgenson, H Coon, M Leppert, N Risch, D Drayna, A Kong, NJ Cox, C Li, M Boehnke, ER Martin, SA Monks, LL Warren, NL Kaplan, MS McPeek, KL Mohlke, MR Erdos, LJ Scott, TE Fingerlin, AU Jackson, K Silander, P Hollstein, M Boehnke, FS Collins, A Oliphant, DL Barker, JR Stuelpnagel, MS Chee, M Olivier, LM Chuang, MS Chang, YT Chen, D Pei, K Ranade, A de Witte, J Allen, N Tran, D Curb, R Pratt, H Neefs, M de Arruda Indig, S Law, B Neri, L Wang, DR Cox, N Patil, AJ Berno, DA Hinds, WA Barrett, JM Doshi, CR Hacker, CR Kautzer, DH Lee, C Marjoribanks, DP McDonough, BT Nguyen, MC Norris, JB Sheehan, N Shen, D Stern, RP Stokowski, DJ Thomas, MO Trulson, KR Vyas, KA Frazer, SP Fodor, DR Cox, NJ Risch, N Risch, J Teng, R Sachidanandam, D Weissman, SC Schmidt, JM Kakol, LD Stein, G Marth, S Sherry, Et Al., H Sengul, DE Weeks, E Feingold, PC Sham, S Purcell, SS Cherny, GR Abecasis, SL Slager, DJ Schaid, RS Spielman, WJ Ewens, AC Syvanen, T Valle, J Tuomilehto, RN Bergman, S Ghosh, ER Hauser, J Eriksson, SJ Nylund, K Kohtamaki, L Toivanen, G Vidgren, E Tuomilehto-Wolf, C Ehnholm, J Blaschak, CD Langefeld, RM Watanabe, V Magnuson, DS Ally, WA Hagopian, E Ross, TA Buchanan, F Collins, M Boehnke, P Van Eerdewegh, RD Little, J Dupuis, RG Del Mastro, K Falls, J Simon, D Torrey, Et Al., AS Whittemore, J Halpern, JK Wolford, D Blunt, C Ballecer, and M Prochazka. Increasing the Power and Efficiency of Disease-Marker Case-Control Association Studies through Use of Allele-Sharing Information. *The American Journal of Human Genetics*, 74(3):432–443, 2002. doi: 10.1086/381652.

Andre Franke, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A Anderson, Joshua C Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G Mathew, Grant W Montgomery, Natalie J Prescott, Soumya Raychaudhuri, Jerome I Rotter, Philip Schumm, Yashoda Sharma, Lisa A Simms, Kent D Tay-

lor, David Whiteman, Cisca Wijmenga, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Büning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jürgen Glas, Andre Van Gossum, Stephen L Guthery, Jonas Halfvarson, Hein W Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julián Panés, Anne Phillips, Deborah D Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A Hillary Steinhart, Pieter C F Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R Targan, Steven R Brant, John D Rioux, Mauro D'Amato, Rinse K Weersma, Subra Kugathasan, Anne M Griffiths, John C Mansfield, Severine Vermeire, Richard H Duerr, Mark S Silverberg, Jack Satsangi, Stefan Schreiber, Judy H Cho, Vito Annese, Hakon Hakonarson, Mark J Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–25, dec 2010. ISSN 1546-1718. doi: 10.1038/ng.717. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299551{&}tool=pmcentrez{&}rendertype=abstract`.

Brenda Gerull, Arnd Heuser, Thomas Wichter, Matthias Paul, Craig T Basson, Deborah A McDermott, Bruce B Lerman, Steve M Markowitz, Patrick T Ellinor, Calum A MacRae, Stefan Peters, Katja S Grossmann, Jörg Drenckhahn, Beate Michely, Sabine Sasse-Klaassen, Walter Birchmeier, Rainer Dietz, Günter Breithardt, Eric Schulze-Bahr, and Ludwig Thierfelder. Mutations in the desmosomal protein plakophilin-2 are common in arrhythmogenic right ventricular cardiomyopathy. *Nature genetics*, 36(11): 1162–4, nov 2004. ISSN 1061-4036. doi: 10.1038/ng1461. URL `http://dx.doi.org/10.1038/ng1461`.

T D Gilmore. Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene*, 25(51):6680–4, oct 2006. ISSN 0950-9232. doi: 10.1038/sj.onc.1209954. URL `http://www.ncbi.nlm.nih.gov/pubmed/17072321`.

B Gilmour, A; Thomson, R; Cullis. Average Information REML: An efficient algorithm for variance estima-

tion in linear mixed models. *Biometrics*, 20:1440–1450, 1995.

Elizabeth Goldmuntz, Prasuna Paluru, Joseph Glessner, Hakon Hakonarson, Jaclyn A Biegel, Peter S White, Xiaowu Gai, and Tamim H Shaikh. Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. *Congenital heart disease*, 6(6):592–602, jan 2011. ISSN 1747-0803. doi: 10.1111/j.1747-0803.2011.00582.x. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4575121{&}tool=pmcentrez{&}rendertype=abstract`.

Runa M Grimholt, Petter Urdal, Olav Klingenberg, and Armin P Piehler. Rapid and reliable detection of α-globin copy number variations by quantitative real-time PCR. *BMC hematology*, 14(1):4, jan 2014. ISSN 2052-1839. doi: 10.1186/2052-1839-14-4. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3904007{&}tool=pmcentrez{&}rendertype=abstract`.

S. Guha. Bayesian Hidden Markon Modelling of Array CGH Data. *Journal of the American Statistical Association*, 103(482):485–497, 2008. doi: 10.1198/016214507000000923.Bayesian.

B J Hayes, P M Visscher, and M E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47–60, feb 2009. ISSN 1469-5073. doi: 10.1017/S0016672308009981. URL `http://www.ncbi.nlm.nih.gov/pubmed/19220931`.

Daniel S Herman, Lien Lam, Matthew R G Taylor, Libin Wang, Polakit Teekakirikul, Danos Christodoulou, Lauren Conner, Steven R DePalma, Barbara McDonough, Elizabeth Sparks, Debbie Lin Teodorescu, Allison L Cirino, Nicholas R Banner, Dudley J Pennell, Sharon Graw, Marco Merlo, Andrea Di Lenarda, Gianfranco Sinagra, J Martijn Bos, Michael J Ackerman, Richard N Mitchell, Charles E Murry, Neal K Lakdawala, Carolyn Y Ho, Paul J R Barton, Stuart a Cook, Luisa Mestroni, J G Seidman, and Christine E Seidman. Truncations of titin causing dilated cardiomyopathy. *The New England journal of medicine*, 366 (7):619–28, feb 2012. ISSN 1533-4406. doi: 10.1056/NEJMoa1110186. URL `http://www.ncbi.nlm.nih.gov/pubmed/22335739`.

Kathleen T Hickey and Kevin Rezzadeh. Hypertrophic cardiomyopathy: a clinical and genetic update. *The*

Nurse practitioner, 38(5):22–31; quiz 31–2, may 2013. ISSN 1538-8662. doi: 10.1097/01.NPR.0000428814.64880.f2. URL http://www.ncbi.nlm.nih.gov/pubmed/23559161.

Carolyn Y Ho. Hypertrophic cardiomyopathy in 2012. *Circulation*, 125(11):1432–8, mar 2012. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.110.017277.

Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, 8(10):e75707, jan 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0075707. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3810480{&}tool=pmcentrez{&}rendertype=abstract.

Simon Hughes, Nona Arneson, Susan Done, and Jeremy Squire. The use of whole genome amplification in the study of human disease. *Progress in biophysics and molecular biology*, 88(1):173–89, may 2005. ISSN 0079-6107. doi: 10.1016/j.pbiomolbio.2004.01.007. URL http://www.ncbi.nlm.nih.gov/pubmed/15561304.

J P Hugot, P Laurent-Puig, C Gower-Rousseau, J M Olson, J C Lee, L Beaugerie, I Naom, J L Dupas, A Van Gossum, M Orholm, C Bonaiti-Pellie, J Weissenbach, C G Mathew, J E Lennard-Jones, A Cortot, J F Colombel, and G Thomas. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature*, 379(6568):821–3, feb 1996. ISSN 0028-0836. doi: 10.1038/379821a0. URL http://www.ncbi.nlm.nih.gov/pubmed/8587604.

J P Hugot, M Chamaillard, H Zouali, S Lesage, J P Cézard, J Belaiche, S Almer, C Tysk, C A O'Morain, M Gassull, V Binder, Y Finkel, A Cortot, R Modigliani, P Laurent-Puig, C Gower-Rousseau, J Macry, J F Colombel, M Sahbatou, and G Thomas. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411(6837):599–603, may 2001. ISSN 0028-0836. doi: 10.1038/35079107. URL http://www.ncbi.nlm.nih.gov/pubmed/11385576.

Y H Jeon, Y-S Heo, C M Kim, Y-L Hyun, T G Lee, S Ro, and J M Cho. Phosphodiesterase: overview of protein structures, potential therapeutic applications and recent progress in drug development. *Cellular and molecular life sciences : CMLS*, 62(11):1198–220, jun 2005. ISSN 1420-682X. doi: 10.1007/s00018-005-4533-5. URL http://www.ncbi.nlm.nih.gov/pubmed/15798894.

Jianming Jiang, Hiroko Wakimoto, J G Seidman, and Christine E Seidman. Allele-specific silencing of mutant Myh6 transcripts in mice suppresses hypertrophic cardiomyopathy. *Science (New York, N.Y.)*, 342(6154):111–4, oct 2013. ISSN 1095-9203. doi: 10.1126/science.1236921. URL http://www.ncbi.nlm.nih.gov/pubmed/24092743.

Takahiro Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering*, 96(4):317–23, jan 2003. ISSN 1389-1723. doi: 10.1016/S1389-1723(03)90130-7. URL http://www.ncbi.nlm.nih.gov/pubmed/16233530.

Hyun Min Kang, Noah a Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23, mar 2008. ISSN 0016-6731. doi: 10.1534/genetics.107.080101. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2278096{&}tool=pmcentrez{&}rendertype=abstract.

Emre Karakoc, Can Alkan, Brian J O'Roak, Megan Y Dennis, Laura Vives, Kenneth Mark, Mark J Rieder, Debbie a Nickerson, and Evan E Eichler. Detection of structural variants and indels within exome data. *Nature methods*, 9(2):176–8, feb 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1810. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3269549{&}tool=pmcentrez{&}rendertype=abstract.

Kasper Karlsson, Ellika Sahlin, Erik Iwarsson, Magnus Westgren, Magnus Nordenskjöld, and Sten Linnarsson. Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. *Genomics*, 105(3):150–8, mar 2015. ISSN 1089-8646. doi: 10.1016/j.ygeno.2014.12.005. URL http://www.sciencedirect.com/science/article/pii/S088875431400278X.

Eimear E Kenny, Itsik Pe'er, Amir Karban, Laurie Ozelius, Adele A Mitchell, Sok Meng Ng, Monica Erazo, Harry Ostrer, Clara Abraham, Maria T Abreu, Gil Atzmon, Nir Barzilai, Steven R Brant, Susan Bressman, Edward R Burns, Yehuda Chowers, Lorraine N Clark, Ariel Darvasi, Dana Doheny, Richard H Duerr, Rami Eliakim, Nir Giladi, Peter K Gregersen, Hakon Hakonarson, Michelle R Jones, Karen Marder, Dermot P B

McGovern, Jennifer Mulle, Avi Orr-Urtreger, Deborah D Proctor, Ann Pulver, Jerome I Rotter, Mark S Silverberg, Thomas Ullman, Stephen T Warren, Matti Waterman, Wei Zhang, Aviv Bergman, Lloyd Mayer, Seymour Katz, Robert J Desnick, Judy H Cho, and Inga Peter. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS genetics*, 8(3):e1002559, 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002559. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3297573{&}tool=pmcentrez{&}rendertype=abstract`.

Eun Kyoung Kim, Sang-Chol Lee, Ji Won Hwang, Sung-A Chang, Sung-Ji Park, Young Keun On, Kyoung Min Park, Yeon Hyeon Choe, Sung-Mok Kim, Seung Woo Park, and Jae K Oh. Differences in apical and non-apical types of hypertrophic cardiomyopathy: a prospective analysis of clinical, echocardiographic, and cardiac magnetic resonance findings and outcome from 350 patients. *European heart journal cardiovascular Imaging*, pages jev192–, aug 2015. ISSN 2047-2412. doi: 10.1093/ehjci/jev192. URL `http://ehjcimaging.oxfordjournals.org/content/early/2015/08/04/ehjci.jev192`.

W C Knowler, R C Williams, D J Pettitt, and a G Steinberg. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American journal of human genetics*, 43(4): 520–6, oct 1988. ISSN 0002-9297. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1715499{&}tool=pmcentrez{&}rendertype=abstract`.

Niklas Krumm, Peter H Sudmant, Arthur Ko, Brian J O'Roak, Maika Malig, Bradley P Coe, Aaron R Quinlan, Deborah a Nickerson, and Evan E Eichler. Copy number variation detection and genotyping from exome sequence data. *Genome research*, 22(8):1525–32, aug 2012. ISSN 1549-5469. doi: 10.1101/gr.138115.112. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3409265{&}tool=pmcentrez{&}rendertype=abstract`.

Shinya Kurata, Takahiro Kanagawa, Yukio Magariyama, Kyoko Takatsu, Kazutaka Yamada, Toyokazu Yokomaku, and Yoichi Kamagata. Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Applied and environmental microbiology*, 70(12):7545–9, dec 2004. ISSN 0099-2240. doi:

10.1128/AEM.70.12.7545-7549.2004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=535213{&}tool=pmcentrez{&}rendertype=abstract.

Kazuto Kurohara, Kouji Komatsu, Tomohiro Kurisaki, Aki Masuda, Naoki Irie, Masahide Asano, Katsuko Sudo, Yo-ichi Nabeshima, Yoichiro Iwakura, and Atsuko Sehara-Fujisawa. Essential roles of Meltrin beta (ADAM19) in heart development. *Developmental biology*, 267(1):14–28, mar 2004. ISSN 0012-1606. doi: 10.1016/j.ydbio.2003.10.021. URL http://www.ncbi.nlm.nih.gov/pubmed/14975714.

Abhimanyu Garg Lalitha Subramanyam, Vinaya Simha. Overlapping syndrome with Familial Partial Lipodystrophy, Dunnigan variety and Cardiomyopathy due to Amino-terminal Heterozygous Missense lamin A/C Mutations. *Clinical genetics*, 78(1):66, 2010.

Hugo Y K Lam, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O'Huallachain, Frederick E Dewey, Lukas Habegger, Euan a Ashley, Mark B Gerstein, Atul J Butte, Hanlee P Ji, and Michael Snyder. Performance comparison of whole-genome sequencing platforms. *Nature biotechnology*, 30(1): 78–82, jan 2012. ISSN 1546-1696. doi: 10.1038/nbt.2065. URL http://www.ncbi.nlm.nih.gov/pubmed/22178993.

Seunggeun Lee, Michael C. Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies SUP. *Biostatistics*, 13(4):762–775, 2012. ISSN 14654644. doi: 10.1093/biostatistics/kxs014.

Seunggeung Lee, GonçaloR. Abecasis, Michael Boehnke, and Xihong Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*, 95(1):5–23, jul 2014. ISSN 00029297. doi: 10.1016/j.ajhg.2014.06.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929714002717.

Bingshan Li and Suzanne M Leal. Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. *American journal of human genetics*, 83(3):311–321, 2008a. doi: 10.1016/j.ajhg.2008.06.024.

Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*, 83(3):311–21, sep 2008b. ISSN 1537-6605. doi: 10.1016/j.ajhg.2008.06.024. URL `http://www.ncbi.nlm.nih.gov/pubmed/18691683http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2842185`.

Duanxiang Li, Ana Morales, and Jorge Gonzalez-quintana. Identification of Novel Mutations in RBM20 in Patients with Dilated Cardiomyopathy. *Clinical and Translational Science*, 3(3):90–97, 2010. doi: 10.1111/j.1752-8062.2010.00198.x.Identification.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, aug 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002{&}tool=pmcentrez{&}rendertype=abstract`.

Mingyao Li, Michael Boehnke, Gonçalo R Abecasis, RJ Klein, C Zeiss, EY Chew, J-Y Tsai, RS Sackler, C Haynes, AK Henning, JP SanGiovanni, SM Mane, ST Mayne, MB Bracken, FL Ferris, J Ott, C Barnstable, J Hoh, DM Maraganore, M de Andrade, TG Lesnick, KJ Strain, MJ Farrer, WA Rocca, PVK Pant, KA Frazer, DR Cox, DG Ballinger, International HapMap Consortium, International HapMap Consortium, N Risch, J Teng, N Risch, TE Fingerlin, M Boehnke, GR Abecasis, M Li, M Boehnke, GR Abecasis, N Risch, JMM Howson, BJ Barratt, JA Todd, HJ Cordell, ES Lander, P Green, L Kruglyak, MJ Daly, MP Reeve-Daly, ES Lander, LE Baum, RS Spielman, RE McGinnis, WJ Ewens, C Cannings, EA Thompson, MP Epstein, X Lin, M Boehnke, JA Nelder, R Mead, D Thompson, JS Witte, M Slattery, D Goldgar, JK Wittke-Thompson, A Pluzhnikov, NJ Cox, SG Self, KY Liang, N Risch, JBS Haldane, N Risch, B Devlin, K Roeder, DB Allison, M Heo, NJ Schork, S-L Wong, RC Elston, JL Haines, MA Hauser, S Schmidt, WK Scott, LM Olson, P Gallins, KL Spencer, SY Kwan, M Noureddine, JR Gilbert, N Schnetz-Boutaud, A Agarwal, EA Postel, MA Pericak-Vance, AO Edwards, R Ritter, KJ Abel, A Manning, C Panhuysen, LA Farrer, S Zareparsi, KE Branham, M Li, S Shah, RJ Klein, J Ott, J Hoh, GR Abecasis, A Swaroop,

J Jakobsdottir, YP Conley, DE Weeks, TS Mah, RE Ferrell, MB Gorin, A Rivera, SA Fisher, LG Fritsche, CN Keilhauer, P Lichtner, T Meitigner, and BHF Weber. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *American journal of human genetics*, 78(5): 778–92, may 2006. ISSN 0002-9297. doi: 10.1086/503711. URL `http://www.ncbi.nlm.nih.gov/pubmed/16642434http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1474028`.

Ilena Egle Astrid Li Mura, Barbara Bauce, Andrea Nava, Manuela Fanciulli, Giovanni Vazza, Elisa Mazzotti, Ilaria Rigato, Marzia De Bortoli, Giorgia Beffagna, Alessandra Lorenzon, Martina Calore, Emanuela Dazzo, Carlo Nobile, Maria Luisa Mostacciuolo, Domenico Corrado, Cristina Basso, Luciano Daliento, Gaetano Thiene, and Alessandra Rampazzo. Identification of a PKP2 gene deletion in a family with arrhythmogenic right ventricular cardiomyopathy. *European journal of human genetics : EJHG*, 21(11): 1226–31, nov 2013. ISSN 1476-5438. doi: 10.1038/ejhg.2013.39. URL `http://www.ncbi.nlm.nih.gov/pubmed/23486541`.

Cecilia M Lindgren, Iris M Heid, Joshua C Randall, Claudia Lamina, Valgerdur Steinthorsdottir, Lu Qi, Elizabeth K Speliotes, Gudmar Thorleifsson, Cristen J Willer, Blanca M Herrera, Anne U Jackson, Noha Lim, Paul Scheet, Nicole Soranzo, Najaf Amin, Yurii S Aulchenko, John C Chambers, Alexander Drong, Jian'an Luan, Helen N Lyon, Fernando Rivadeneira, Serena Sanna, Nicholas J Timpson, M Carola Zillikens, Jing Hua Zhao, Peter Almgren, Stefania Bandinelli, Amanda J Bennett, Richard N Bergman, Lori L Bonnycastle, Suzannah J Bumpstead, Stephen J Chanock, Lynn Cherkas, Peter Chines, Lachlan Coin, Cyrus Cooper, Gabriel Crawford, Angela Doering, Anna Dominiczak, Alex S F Doney, Shah Ebrahim, Paul Elliott, Michael R Erdos, Karol Estrada, Luigi Ferrucci, Guido Fischer, Nita G Forouhi, Christian Gieger, Harald Grallert, Christopher J Groves, Scott Grundy, Candace Guiducci, David Hadley, Anders Hamsten, Aki S Havulinna, Albert Hofman, Rolf Holle, John W Holloway, Thomas Illig, Bo Isomaa, Leonie C Jacobs, Karen Jameson, Pekka Jousilahti, Fredrik Karpe, Johanna Kuusisto, Jaana Laitinen, G Mark Lathrop, Debbie A Lawlor, Massimo Mangino, Wendy L McArdle, Thomas Meitinger, Mario A Morken, Andrew P Morris, Patricia Munroe, Narisu Narisu, Anna Nordström, Peter Nordström, Ben A Oostra, Colin N A Palmer, Felicity Payne, John F Peden, Inga Prokopenko, Frida Renström, Aimo Ruokonen,

Veikko Salomaa, Manjinder S Sandhu, Laura J Scott, Angelo Scuteri, Kaisa Silander, Kijoung Song, Xin Yuan, Heather M Stringham, Amy J Swift, Tiinamaija Tuomi, Manuela Uda, Peter Vollenweider, Gerard Waeber, Chris Wallace, G Bragi Walters, Michael N Weedon, Jacqueline C M Witteman, Cuilin Zhang, Weihua Zhang, Mark J Caulfield, Francis S Collins, George Davey Smith, Ian N M Day, Paul W Franks, Andrew T Hattersley, Frank B Hu, Marjo-Riitta Jarvelin, Augustine Kong, Jaspal S Kooner, Markku Laakso, Edward Lakatta, Vincent Mooser, Andrew D Morris, Leena Peltonen, Nilesh J Samani, Timothy D Spector, David P Strachan, Toshiko Tanaka, Jaakko Tuomilehto, André G Uitterlinden, Cornelia M van Duijn, Nicholas J Wareham, Hugh Watkins, Dawn M Waterworth, Michael Boehnke, Panos Deloukas, Leif Groop, David J Hunter, Unnur Thorsteinsdottir, David Schlessinger, H-Erich Wichmann, Timothy M Frayling, Gonçalo R Abecasis, Joel N Hirschhorn, Ruth J F Loos, Kari Stefansson, Karen L Mohlke, Inês Barroso, and Mark I McCarthy. Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS genetics*, 5(6):e1000508, jun 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000508. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2695778{&}tool=pmcentrez{&}rendertype=abstract`.

Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–5, jan 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1681. URL `http://www.ncbi.nlm.nih.gov/pubmed/21892150`.

L.R. Lopes, C. Murphy, P. Syrris, C. Dalageorgou, W.J. McKenna, P.M. Elliott, and V. Plagnol. Use of High-throughput Targeted Exome-sequencing to screen for Copy Number Variation in Hypertrophic Cardiomyopathy. *European Journal of Medical Genetics*, pages 1–6, 2015. ISSN 17697212. doi: 10.1016/j.ejmg.2015.10.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S1769721215300252`.

Luís R Lopes, M Shafiqur Rahman, and Perry M Elliott. A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic cardiomyopathy caused by sarcomeric protein mutations. *Heart (British Cardiac Society)*, 99(24):1800–11, dec 2013a. ISSN 1468-201X. doi: 10.1136/heartjnl-2013-303939. URL `http://www.ncbi.nlm.nih.gov/pubmed/23674365`.

Luis R Lopes, Anna Zekavati, Petros Syrris, Mike Hubank, Claudia Giambartolomei, Chrysoula Dalageorgou, Sharon Jenkins, William McKenna, Vincent Plagnol, and Perry M Elliott. Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *Journal of medical genetics*, 50(4):228–39, apr 2013b. ISSN 1468-6244. doi: 10.1136/jmedgenet-2012-101270. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607113{&}tool=pmcentrez{&}rendertype=abstract`.

Luis R Lopes, Petros Syrris, Oliver P Guttmann, Constantinos O'Mahony, Hak Chiaw Tang, Chrysoula Dalageorgou, Sharon Jenkins, Mike Hubank, Lorenzo Monserrat, William J McKenna, Vincent Plagnol, and Perry M Elliott. Novel genotype-phenotype associations demonstrated by high-throughput sequencing in patients with hypertrophic cardiomyopathy. *Heart (British Cardiac Society)*, pages heartjnl–2014–306387–, oct 2014. ISSN 1468-201X. doi: 10.1136/heartjnl-2014-306387. URL `http://heart.bmj.com/content/early/2014/10/28/heartjnl-2014-306387.long`.

V M Lourenço, a M Pires, and M Kirst. Robust linear regression methods in association studies. *Bioinformatics (Oxford, England)*, 27(6):815–21, mar 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr006.

Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384, mar 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000384. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2633048{&}tool=pmcentrez{&}rendertype=abstract`.

Ali J Marian. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends in cardiovascular medicine*, 22(8):219–23, nov 2012. ISSN 1873-2615. doi: 10.1016/j.tcm.2012.08.001. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3496831{&}tool=pmcentrez{&}rendertype=abstract`.

G McKoy, N Protonotarios, a Crosby, a Tsatsopoulou, a Anastasakis, a Coonar, M Norman, C Baboonian, S Jeffery, and W J McKenna. Identification of a deletion in plakoglobin in arrhythmogenic right ventricular cardiomyopathy with palmoplantar keratoderma and woolly hair (Naxos disease). *Lancet*, 355(9221):2119–

24, jun 2000. ISSN 0140-6736. doi: 10.1016/S0140-6736(00)02379-5. URL http://www.ncbi.nlm.nih.gov/pubmed/10902626.

Janine Meienberg, Katja Zerjavic, Irene Keller, Michal Okoniewski, Andrea Patrignani, Katja Ludin, Zhenyu Xu, Beat Steinmann, Thierry Carrel, Benno R??thlisberger, Ralph Schlapbach, Rcrossedmy Bruggmann, and Gabor Matyas. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Research*, 43(11), 2015. ISSN 13624962. doi: 10.1093/nar/gkv216.

A.; WrightA. Meindl. A gene (RPGR) with homology to the RCC1 guanine nucleotide exchange factor is mutated in Xlinked retinitis pigmentosa (RP3). *Nature ...*, 13:35–43, 1996. URL http://www.nature.com/ng/journal/v13/n1/abs/ng0596-35.html.

P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science (New York, N.Y.)*, 201(4358):786–92, sep 1978. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/356262.

Lynne Martina Millar and Sanjay Sharma. Genetics of sudden Cardiac Death. pages 1–9, 2015. ISSN 1523-3782. doi: 10.1007/s11886-015-0606-8.

G E Moore. Cramming more components onto integrated circuits (Reprinted from Electronics, pg 114-117, April 19, 1965). *Proceedings Of The Ieee*, 86(1):82–85, 1998. ISSN 1098-4232. doi: 10.1109/N-SSC.2006.4785860. URL papers3://publication/uuid/8E5EB7C8-681C-447D-9361-E68D1932997D.

Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research*, 615(1-2):28–56, feb 2007. ISSN 0027-5107. doi: 10.1016/j.mrfmmm.2006.09.003. URL http://www.ncbi.nlm.nih.gov/pubmed/17101154.

Andrew P. Morris, Cecilia M. Lindgren, Eleftheria Zeggini, Nicholas J. Timpson, Timothy M. Frayling, Andrew T. Hattersley, and Mark I. McCarthy. A powerful approach to sub-phenotype analysis in

population-based genetic association studies. *Genetic Epidemiology*, 34:335–343, 2010. ISSN 07410395. doi: 10.1002/gepi.20486.

G L Mutter and K A Boynton. PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic acids research*, 23(8):1411–8, apr 1995. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=306870{&}tool=pmcentrez{&}rendertype=abstract.

Nico J D Nagelkerke, Barbara Hoebee, Peter Teunis, and Tjeerd G Kimman. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *European journal of human genetics : EJHG*, 12(11):964–70, nov 2004. ISSN 1018-4813. doi: 10.1038/sj.ejhg.5201255. URL http://www.ncbi.nlm.nih.gov/pubmed/15340361.

Manuel Neiva-Sousa, João Almeida-Coelho, Inês Falcão-Pires, and Adelino F. Leite-Moreira. Titin mutations: the fall of Goliath. *Heart Failure Reviews*, pages 579–588, 2015. ISSN 1382-4147. doi: 10.1007/s10741-015-9495-6. URL http://link.springer.com/10.1007/s10741-015-9495-6.

Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, Michael Bamshad, Deborah a Nickerson, and Jay Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461 (7261):272–6, sep 2009. ISSN 1476-4687. doi: 10.1038/nature08250. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2844771{&}tool=pmcentrez{&}rendertype=abstract.

Michael Nothnagel, Alexander Herrmann, Andreas Wolf, Stefan Schreiber, Matthias Platzer, Reiner Siebert, Michael Krawczak, and Jochen Hampe. Technology-specific error signatures in the 1000 Genomes Project data. *Human genetics*, 130(4):505–16, oct 2011. ISSN 1432-1203. doi: 10.1007/s00439-011-0971-3. URL http://www.ncbi.nlm.nih.gov/pubmed/21344269.

L M Nunn and P D Lambiase. Genetics and cardiovascular disease–causes and prevention of unexpected sudden adult death: the role of the SADS clinic. *Heart (British Cardiac Society)*, 97(14):1122–1127, 2011. ISSN 1355-6037. doi: 10.1136/hrt.2010.218511.

Laurence M. Nunn, Luis R. Lopes, Petros Syrris, Cian Murphy, Vincent Plagnol, Eileen Firman, Chrysoula Dalageorgou, Esther Zorio, Diana Domingo, Victoria Murday, Iain Findlay, Alexis Duncan, Gerry Carr-White, Leema Robert, Teofila Bueser, Caroline Langman, Simon P Fynn, Martin Goddard, Anne White, Henning Bundgaard, Laura Ferrero-Miliani, Nigel Wheeldon, Simon K. Suvarna, Aliceson O'Beirne, Martin D. Lowe, William J. McKenna, Perry M. Elliott, and Pier D. Lambiase. Diagnostic yield of molecular autopsy in patients with sudden arrhythmic death syndrome using targeted exome sequencing. *Europace*, page euv285, 2015. ISSN 1099-5129. doi: 10.1093/europace/euv285. URL `http://europace.oxfordjournals.org/lookup/doi/10.1093/europace/euv285`.

Laurence M Nunn, Luis R Lopes, Petros Syrris, Cian Murphy, Vincent Plagnol, Eileen Firman, Chrysoula Dalageorgou, Esther Zorio, Diana Domingo, Victoria Murday, Iain Findlay, Alexis Duncan, Gerry Carr-White, Leema Robert, Teofila Bueser, Caroline Langman, Simon P Fynn, Martin Goddard, Anne White, Henning Bundgaard, Laura Ferrero-Miliani, Nigel Wheeldon, Simon K Suvarna, Aliceson O'Beirne, Martin D Lowe, William J McKenna, Perry M Elliott, Pier D Lambiase, E. Behr, A. Casey, M. Sheppard, M. Wright, T. Bowker, M. Davies, SG. Priori, AA. Wilde, M. Horie, Y. Cho, ER. Behr, C. Berul, LM. Nunn, PD. Lambiase, R. Bagnall, J. Das, J. Duflou, C. Semsarian, TG. Consortium, I. Splawski, J. Shen, K. Timothy, M. Lehmann, S. Priori, J. Robinson, N. Kambouris, H. Nuss, D. Johns, G. Tomaselli, E. Marban, J. Balser, D. Tester, M. Will, C. Haglund, M. Ackerman, J. Kapplinger, D. Tester, M. Alders, B. Benito, M. Berthet, J. Brugada, I. Goldenberg, S. Horr, A. Moss, C. Lopes, A. Barsheshet, S. McNitt, M. Kawamura, S. Ohno, N. Naiki, I. Nagaoka, K. Dochi, Q. Wang, V. Fressart, G. Duthoit, E. Donal, V. Probst, J. Deharo, P. Chevalier, E. Burashnikov, R. Pfeiffer, H. Barajas-Martinez, E. Delpon, D. Hu, M. Desai, L. Crotti, C. Marcou, D. Tester, S. Castelletti, J. Giudicessi, M. Torchio, S. Kapa, D. Tester, B. Salisbury, C. Harris-Kerr, M. Pungliya, M. Alders, L. Refsgaard, A. Holst, G. Sadjadieh, S. Haunsø, J. Nielsen, M. Olesen, MS. Olesen, NF. Jensen, AG. Holst, JB. Nielsen, J. Tfelt-Hansen, T. Jespersen, B. Risgaard, R. Jabbari, L. Refsgaard, A. Holst, S. Haunsø, A. Sadjadieh, P. Mohler, I. Splawski, C. Napolitano, G. Bottelli, L. Sharpe, K. Timothy, P. Mohler, S. Le Scouarnec, I. Denjoy, J. Lowe, P. Guicheney, L. Caron, J. Sherman, D. Tester, M. Ackerman, P. Garcia-Pavia, P. Syrris, C. Salas, A. Evans, J. Mirelis,

M. Cobo-Marcos, J. Genschel, B. Bochow, S. Kuepferling, R. Ewert, R. Hetzer, H. Lochs, J. Genschel, H. Schmidt, J. Rankin, M. Auer-Grumbach, W. Bagg, K. Colclough, T. Nguyen, J. Fenton-May, E. Purevjav, T. Arimura, S. Augustin, A. Huby, K. Takagi, S. Nunoda, M. Refaat, S. Lubitz, S. Makino, Z. Islam, J. Frangiskakis, H. Mehdi, J. Gomes, M. Finlay, AK. Ahmed, EJ. Ciaccio, A. Asimaki, JE. Saffitz, NE. Hasselberg, T. Edvardsen, H. Petri, KE. Berge, TP. Leren, H. Bundgaard, J. Punetha, EP. Hoffman, DS. Herman, L. Lam, MRG. Taylor, L. Wang, P. Teekakirikul, D. Christodoulou, C. Andreasen, JB. Nielsen, L. Refsgaard, AG. Holst, AH. Christensen, L. Andreasen, O. Campuzano, C. Allegue, A. Fernandez, A. Iglesias, R. Brugada, P. Postema, I. Christiaans, N. Hofman, M. Alders, T. Koopmann, and C. Bezzina. Diagnostic yield of molecular autopsy in patients with sudden arrhythmic death syndrome using targeted exome sequencing. *Europace*, 18(6):888–96, jun 2016. ISSN 1532-2092. doi: 10.1093/europace/euv285. URL `http://www.ncbi.nlm.nih.gov/pubmed/26498160`.

Y Ogura, D K Bonen, N Inohara, D L Nicolae, F F Chen, R Ramos, H Britton, T Moran, R Karaliuskas, R H Duerr, J P Achkar, S R Brant, T M Bayless, B S Kirschner, S B Hanauer, G Nuñez, and J H Cho. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603–6, may 2001. ISSN 0028-0836. doi: 10.1038/35079114. URL `http://www.ncbi.nlm.nih.gov/pubmed/11385577`.

a E Oostlander, G a Meijer, and B Ylstra. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clinical genetics*, 66(6):488–95, dec 2004. ISSN 0009-9163. doi: 10.1111/j.1399-0004.2004.00322.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/15521975`.

Alistair T Pagnamenta, Richard Holt, Mohammed Yusuf, Dalila Pinto, Kirsty Wing, Catalina Betancur, Stephen W Scherer, Emanuela V Volpi, and Anthony P Monaco. A family with autism and rare copy number variants disrupting the Duchenne/Becker muscular dystrophy gene DMD and TRPM3. *Journal of neurodevelopmental disorders*, 3(2):124–31, jun 2011. ISSN 1866-1955. doi: 10.1007/s11689-011-9076-5. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3105230{&}tool=pmcentrez{&}rendertype=abstract`.

Cameron Palmer and Itsik Pe'er. Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. *PLOS Genetics*, 12(6):e1006091, 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006091. URL `http://dx.plos.org/10.1371/journal.pgen.1006091`.

J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1114759108.

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, dec 2006. ISSN 1553-7404. doi: 10.1371/journal.pgen.0020190. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1713260{&}tool=pmcentrez{&}rendertype=abstract`.

Alexandra Pérez-Serra, Rocío Toro, Oscar Campuzano, Georgia Sarquella-Brugada, Paola Berne, Anna Iglesias, Alipio Mangas, Josep Brugada, and Ramon Brugada. A Novel Mutation in Lamin A/C Causing Familial Dilated Cardiomyopathy Associated With Sudden Cardiac Death. *Journal of Cardiac Failure*, 21 (3):217–225, 2015. ISSN 10719164. doi: 10.1016/j.cardfail.2014.12.003.

Robert Pinard, Alex de Winter, Gary J Sarkis, Mark B Gerstein, Karrie R Tartaro, Ramona N Plant, Michael Egholm, Jonathan M Rothberg, and John H Leamon. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics*, 7: 216, jan 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-216. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560136{&}tool=pmcentrez{&}rendertype=abstract`.

Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, Dinakantha Kumararatne, Rainer Doffinger, and Sergey Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics (Oxford, England)*, 28(21):2747–54,

nov 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts526. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3476336{&}tool=pmcentrez{&}rendertype=abstract.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–9, aug 2006. ISSN 1061-4036. doi: 10.1038/ng1847. URL http://dx.doi.org/10.1038/ng1847.

J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, jun 2000. ISSN 0016-6731. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461096{&}tool=pmcentrez{&}rendertype=abstract.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel a R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, sep 2007. ISSN 0002-9297. doi: 10.1086/519795. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838{&}tool=pmcentrez{&}rendertype=abstract.

Fedik Rahimov and Louis M Kunkel. The cell biology of disease: cellular and molecular mechanisms underlying muscular dystrophy. *The Journal of cell biology*, 201(4):499–510, may 2013. ISSN 1540-8140. doi: 10.1083/jcb.201212142. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3653356{&}tool=pmcentrez{&}rendertype=abstract.

Alessandra Rampazzo, Andrea Nava, Sandro Malacrida, Giorgia Beffagna, Barbara Bauce, Valeria Rossi, Rosanna Zimbello, Barbara Simionati, Cristina Basso, Gaetano Thiene, Jeffrey a Towbin, and Gian a Danieli. Mutation in human desmoplakin domain binding to plakoglobin causes a dominant form of arrhythmogenic right ventricular cardiomyopathy. *American journal of human genetics*, 71(5):1200–6, nov 2002. ISSN 0002-9297. doi: 10.1086/344208. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=385098{&}tool=pmcentrez{&}rendertype=abstract.

Elliott Rees, James T R Walters, Lyudmila Georgieva, Anthony R Isles, Kimberly D Chambert, Alexander L Richards, Gerwyn Mahoney-Davies, Sophie E Legge, Jennifer L Moran, Steven a McCarroll, Michael C O'Donovan, Michael J Owen, and George Kirov. Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry : the journal of mental science*, 204:108–14, feb 2014. ISSN 1472-1465. doi: 10.1192/bjp.bp.113.131052. URL http://www.ncbi.nlm.nih.gov/pubmed/24311552.

Nora Rieber, Marc Zapatka, Bärbel Lasitschka, David Jones, Paul Northcott, Barbara Hutter, Natalie Jäger, Marcel Kool, Michael Taylor, Peter Lichter, Stefan Pfister, Stephan Wolf, Benedikt Brors, and Roland Eils. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PloS one*, 8(6):e66621, jan 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0066621. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679043{&}tool=pmcentrez{&}rendertype=abstract.

N Risch and J Teng. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome research*, 8(12):1273–88, dec 1998. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/9872982.

N J Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–56, jun 2000. ISSN 0028-0836. doi: 10.1038/35015718. URL http://www.ncbi.nlm.nih.gov/pubmed/10866211.

J D Roberts, J C Herkert, J Rutberg, S M Nikkel, a C P Wiesfeld, D Dooijes, R M Gow, J P van Tintelen, and M H Gollob. Detection of genomic deletions of PKP2 in arrhythmogenic right ventricular cardiomyopathy. *Clinical genetics*, 83(5):452–6, may 2013. ISSN 1399-0004. doi: 10.1111/j.1399-0004.2012.01950.x. URL http://www.ncbi.nlm.nih.gov/pubmed/22889254.

Jorge Romero, Eliany Mejia-Lopez, Carlos Manrique, and Richard Lucariello. Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC/D): A Systematic Literature Review. *Clinical Medicine Insights. Cardiology*, 7:97–114, jan 2013. ISSN 1179-5468. doi: 10.4137/CMC.S10940. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3667685{&}tool=pmcentrez{&}rendertype=abstract.

V Rossi, I Bally, and M Lacroix. Classical Complement Pathway Components C1r and C1s: Purification from Human Serum and in Recombinant Form and Functional Characterization. *The Complement ...*, 2014. URL http://books.google.com/books?hl=en{&}lr={&}id=UUMHh9tjSE0C{&}oi=fnd{&}pg=PR13{&}dq=The+Complement+System{&}ots=lXaVfdDUaN{&}sig=fUskVypAqmUyscWLYqQkcBabaMMhttp://link.springer.com/protocol/10.1007/978-1-62703-724-2{_}4.

Srijita Sen-Chowdhry, Petros Syrris, and William J McKenna. Role of genetic analysis in the management of patients with arrhythmogenic right ventricular dysplasia/cardiomyopathy. *Journal of the American College of Cardiology*, 50(19):1813–21, nov 2007. ISSN 1558-3597. doi: 10.1016/j.jacc.2007.08.008. URL http://www.ncbi.nlm.nih.gov/pubmed/17980246.

Panagiotis I Sergouniotis, Christina Chakarova, Cian Murphy, Mirjana Becker, Eva Lenassi, Gavin Arno, Monkol Lek, Daniel G Macarthur, Shomi S Bhattacharya, Anthony T Moore, Graham E Holder, Anthony G Robson, Uwe Wolfrum, Andrew R Webster, and Vincent Plagnol. AJHG The American Journal of Human Genetics Biallelic variants in TTLL5 , encoding a tubulin glutamylase , cause retinal dystrophy. *American journal of human genetics*, 94(5):760–9, 2014.

Udai P Singh, Angela E Murphy, Reilly T Enos, Haidar A Shamran, Narendra P Singh, Honbing Guan, Venkatesh L Hegde, Daping Fan, Robert L Price, Dennis D Taub, Manoj K Mishra, Mitzi Nagarkatti, and Prakash S Nagarkatti. miR-155 deficiency protects mice from experimental colitis by reducing T helper type 1/type 17 responses. *Immunology*, 143(3):478–89, nov 2014. ISSN 1365-2567. doi: 10.1111/imm.12328. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4212960{&}tool=pmcentrez{&}rendertype=abstract.

R S Spielman, R E McGinnis, and W J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American journal of human genetics*, 52 (3):506–16, mar 1993. ISSN 0002-9297. URL http://www.ncbi.nlm.nih.gov/pubmed/8447318http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1682161.

Petros Syrris, Deirdre Ward, Angeliki Asimaki, Alison Evans, Srijita Sen-Chowdhry, Sian E Hughes,

and William J McKenna. Desmoglein-2 mutations in arrhythmogenic right ventricular cardiomyopathy: a genotype-phenotype characterization of familial disease. *European heart journal*, 28(5):581–8, mar 2007. ISSN 0195-668X. doi: 10.1093/eurheartj/ehl380. URL http://www.ncbi.nlm.nih.gov/pubmed/17105751.

Renjie Tan, Yadong Wang, Sarah E. Kleinstein, Yongzhuang Liu, Xiaolin Zhu, Hongzhe Guo, Qinghua Jiang, Andrew S. Allen, and Mingfu Zhu. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Human Mutation*, 35(7):899–907, 2014. ISSN 10981004. doi: 10.1002/humu.22537.

Matthew Taylor, Sharon Graw, Gianfranco Sinagra, Carl Barnes, Dobromir Slavov, Francesca Brun, Bruno Pinamonti, Ernesto E Salcedo, William Sauer, Stylianos Pyxaras, Brian Anderson, Bernd Simon, Julius Bogomolovas, Siegfried Labeit, Henk Granzier, and Luisa Mestroni. Genetic variation in titin in arrhythmogenic right ventricular cardiomyopathy-overlap syndromes. *Circulation*, 124(8):876–85, aug 2011. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.110.005405. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3167235{&}tool=pmcentrez{&}rendertype=abstract.

Jamie K Teer and James C Mullikin. Exome sequencing: the sweet spot before whole genomes. *Human molecular genetics*, 19(R2):R145–51, oct 2010. ISSN 1460-2083. URL http://hmg.oxfordjournals.org/cgi/content/abstract/19/R2/R145.

Francesco Testa, Settimio Rossi, Raffaella Colucci, Beatrice Gallo, Valentina Di Iorio, Michele Della Corte, Claudio Azzolini, Paolo Melillo, and Francesca Simonelli. Macular abnormalities in Italian patients with retinitis pigmentosa. *The British journal of ophthalmology*, feb 2014. ISSN 1468-2079. doi: 10.1136/bjophthalmol-2013-304082. URL http://www.ncbi.nlm.nih.gov/pubmed/24532797.

Gaetano Thiene, Cristina Basso, Gianantonio Danieli, Alessandra Rampazzo, Domenico Corrado, and Andrea Nava. Right Ventricular Cardiomyopathy A Still Underrecognized Clinic Entity. *Trends in cardiovascular medicine*, 7(3), 1997.

N Tiso, D a Stephan, a Nava, a Bagattin, J M Devaney, F Stanchi, G Larderet, B Brahmbhatt, K Brown, B Bauce, M Muriago, C Basso, G Thiene, G a Danieli, and a Rampazzo. Identification of mutations in the cardiac ryanodine receptor gene in families affected with arrhythmogenic right ventricular cardiomyopathy type 2 (ARVD2). *Human molecular genetics*, 10(3):189–94, feb 2001. ISSN 0964-6906. URL http://www.ncbi.nlm.nih.gov/pubmed/11159936.

John A Todd, Neil M Walker, Jason D Cooper, Deborah J Smyth, Kate Downes, Vincent Plagnol, Rebecca Bailey, Sergey Nejentsev, Sarah F Field, Felicity Payne, Christopher E Lowe, Jeffrey S Szeszko, Jason P Hafler, Lauren Zeitels, Jennie H M Yang, Adrian Vella, Sarah Nutland, Helen E Stevens, Helen Schuilenburg, Gillian Coleman, Meeta Maisuria, William Meadows, Luc J Smink, Barry Healy, Oliver S Burren, Alex A C Lam, Nigel R Ovington, James Allen, Ellen Adlem, Hin-Tak Leung, Chris Wallace, Joanna M M Howson, Cristian Guja, Constantin Ionescu-Tîrgovite, Matthew J Simmonds, Joanne M Heward, Stephen C L Gough, David B Dunger, Linda S Wicker, and David G Clayton. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, 39(7): 857–64, jul 2007. ISSN 1061-4036. doi: 10.1038/ng2068. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2492393{&}tool=pmcentrez{&}rendertype=abstract.

Zhongyi Tong, Bimei Jiang, Yanyang Wu, Yanjuan Liu, Yuanbin Li, Min Gao, Yu Jiang, Qinglan Lv, and Xianzhong Xiao. MiR-21 Protected Cardiomyocytes against Doxorubicin-Induced Apoptosis by Targeting BTG2. *International journal of molecular sciences*, 16(7):14511–25, jan 2015. ISSN 1422-0067. doi: 10.3390/ijms160714511. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4519855{&}tool=pmcentrez{&}rendertype=abstract.

Benjamin Tournier, Caroline Chapusot, Emilie Courcet, Laurent Martin, Côme Lepage, Jean Faivre, and Françoise Piard. Why do results conflict regarding the prognostic value of the methylation status in colon cancers? the role of the preservation method. *BMC Cancer*, 12(1):12, 2012. ISSN 1471-2407. doi: 10.1186/1471-2407-12-12. URL http://www.biomedcentral.com/1471-2407/12/12.

Gerrida M Uys, Amsha Ramburan, Benjamin Loos, Craig J Kinnear, Lundi J Korkie, Jomien Mouton,

Johann Riedemann, and Johanna C Moolman-Smook. Myomegalin is a novel A-kinase anchoring protein involved in the phosphorylation of cardiac myosin binding protein C. *BMC cell biology*, 12(1):18, jan 2011. ISSN 1471-2121. doi: 10.1186/1471-2121-12-18. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3103437{&}tool=pmcentrez{&}rendertype=abstract.

Tatiyana Vaikhanskaya, Larysa Sivitskaya, Nina Danilenko, Oleg Davydenko, Tatsiyana Kurushka, and Irina Sidorenko. LMNA-related dilated cardiomyopathy. *Oxford medical case reports*, 2014(6):102–4, sep 2014. ISSN 2053-8855. doi: 10.1093/omcr/omu040. URL http://www.ncbi.nlm.nih.gov/pubmed/25988045http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4369987.

Hermine a van Duyvenvoorde, Julian C Lui, Sarina G Kant, Wilma Oostdijk, Antoinet Cj Gijsbers, Mariëtte Jv Hoffer, Marcel Karperien, Marie Je Walenkamp, Cees Noordam, Paul G Voorhoeve, Verónica Mericq, Alberto M Pereira, Hedi L Claahsen-van de Grinten, Sandy a van Gool, Martijn H Breuning, Monique Losekoot, Jeffrey Baron, Claudia Al Ruivenkamp, and Jan M Wit. Copy number variants in patients with short stature. *European journal of human genetics : EJHG*, (April):1–8, sep 2013. ISSN 1476-5438. doi: 10.1038/ejhg.2013.203. URL http://www.ncbi.nlm.nih.gov/pubmed/24065112.

Higuchi R. Walsh PS, Erlich HA. Preferential PCR amplification of alleles: mechanisms and solutions. - PubMed - NCBI, 1992. URL http://www.ncbi.nlm.nih.gov/pubmed/?term=Preferential+PCR+amplifi+cation+of+alleles{%}3A+mechanisms+and+solutions.

Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, sep 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq603. URL http://nar.oxfordjournals.org/content/38/16/e164.

Yun Wang, John J Digiovanna, Jere B Stern, Thomas J Hornyak, Mark Raffeld, Sikandar G Khan, Kyu-Seon Oh, M Christine Hollander, Philip a Dennis, and Kenneth H Kraemer. Evidence of ultraviolet type mutations in xeroderma pigmentosum melanomas. *Proceedings of the National Academy of Sciences of the United States of America*, 106(15):6279–6284, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0812401106.

Zuoheng Wang, Xiangtao Liu, Bao-Zhu Yang, and Joel Gelernter. The role and challenges of exome sequencing in studies of human diseases. *Frontiers in genetics*, 4(August):160, jan 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00160. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3752524{&}tool=pmcentrez{&}rendertype=abstract.

Ke Wei, Vahid Serpooshan, Cecilia Hurtado, Marta Diez-Cuñado, Mingming Zhao, Sonomi Maruyama, Wenhong Zhu, Giovanni Fajardo, Michela Noseda, Kazuto Nakamura, Xueying Tian, Qiaozhen Liu, Andrew Wang, Yuka Matsuura, Paul Bushway, Wenqing Cai, Alex Savchenko, Morteza Mahmoudi, Michael D Schneider, Maurice J B van den Hoff, Manish J Butte, Phillip C Yang, Kenneth Walsh, Bin Zhou, Daniel Bernstein, Mark Mercola, and Pilar Ruiz-Lozano. Epicardial FSTL1 reconstitution regenerates the adult mammalian heart. *Nature*, 525(7570):479–85, sep 2015. ISSN 1476-4687. doi: 10.1038/nature15372. URL http://www.ncbi.nlm.nih.gov/pubmed/26375005.

Shu-Hui Wen and Miao-Yu Tsai. Haplotype association analysis of combining unrelated case-control and triads with consideration of population stratification. *Frontiers in Genetics*, 5:103, apr 2014. ISSN 1664-8021. doi: 10.3389/fgene.2014.00103. URL http://journal.frontiersin.org/article/10.3389/fgene.2014.00103/abstract.

Yalu Wen and Qing Lu. A multiclass likelihood ratio approach for genetic risk prediction allowing for phenotypic heterogeneity. *Genetic epidemiology*, 37(7):715–25, nov 2013. ISSN 1098-2272. doi: 10.1002/gepi.21751. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3917316{&}tool=pmcentrez{&}rendertype=abstract.

Tanja Woyke, Damon Tighe, Konstantinos Mavromatis, Alicia Clum, Alex Copeland, Wendy Schackwitz, Alla Lapidus, Dongying Wu, John P McCutcheon, Bradon R McDonald, Nancy A Moran, James Bristow, and Jan-Fang Cheng. One bacterial cell, one complete genome. *PloS one*, 5(4):e10314, jan 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010314. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859065{&}tool=pmcentrez{&}rendertype=abstract.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-

variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, 89(1):82–93, jul 2011. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.05.029. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3135811{&}tool=pmcentrez{&}rendertype=abstract.

Saranya P. Wyles, Xing Li, Sybil C. Hrstka, Santiago Reyes, Saji Oommen, Rosanna Beraldi, Jessica Edwards, Andre Terzic, Timothy M. Olson, and Timothy J. Nelson. Modeling structural and functional deficiencies of RBM20 familial dilated cardiomyopathy using human induced pluripotent stem cells. *Human Molecular Genetics*, (November):ddv468, 2015. ISSN 0964-6906. doi: 10.1093/hmg/ddv468. URL http://www.hmg.oxfordjournals.org/lookup/doi/10.1093/hmg/ddv468.

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006. ISSN 1061-4036. doi: 10.1038/ng1702.

Bruno Zeitouni, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-né, Alain Nicolas, Olivier Delattre, and Emmanuel Barillot. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)*, 26(15):1895–6, aug 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq293. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2905550{&}tool=pmcentrez{&}rendertype=abstract.

Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael a Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, and Edward S Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–60, apr 2010. ISSN 1546-1718. doi: 10.1038/ng.546. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2931336{&}tool=pmcentrez{&}rendertype=abstract.

Wanding Zhou, Tenghui Chen, Hao Zhao, Agda Karina Eterovic, Funda Meric-Bernstam, Gordon B. Mills,

and Ken Chen. Bias from removing read duplication in ultra-deep sequencing experiments. *Bioinformatics*, 30(8):1073–1080, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btt771.

D. P. Zipes and H. J. J. Wellens. Sudden Cardiac Death. *Circulation*, 98(21):2334–2351, nov 1998. ISSN 0009-7322. doi: 10.1161/01.CIR.98.21.2334. URL `http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.98.21.2334`.