The genetic architecture of hypertrophic cardiomyopathy

Dr. Andrew R. Harper

Hertford College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2020

Abstract

Hypertrophic cardiomyopathy (HCM) is the most common genetic heart disorder, affecting at least 1 in 500 individuals, and a leading cause of sudden death. Genetic testing for rare, causal, genetic variants in sarcomere genes is the standard of care and conducted at scale. However, more than half of HCM patients do not carry identifiable pathogenic variants and, in those that do, there is substantial variation in penetrance and disease expression.

Here, the genetic architecture of HCM is further evaluated, under a central hypothesis that the genetic aetiology of HCM extends beyond known rare variant contributions. Through a series of case-control analyses monogenic, oligogenic and polygenic models of disease were assessed. Burden testing analyses support prior knowledge regarding the monogenic basis to HCM. Quantitative analyses directed towards quantifying the penetrance and expressivity of disease-causing HCM variants were largely underpowered. Similarly, systematic evaluation for oligogenicity was underpowered. However, haplotype analysis of a candidate variant (*MYBPC3*Δ25) presumed to be of importance to oligogenicity revealed synthetic association with a rare pathogenic variant (*MYBPC3* c.1224-52G>A), quelling this specific oligogenic hypothesis. Polygenicity was evaluated through genome wide association analyses. The additive effects of common variants explained 34.0$\pm$2.4% of phenotypic variance in sarcomere-negative HCM, and 15.8$\pm$3.8% in sarcomere-positive HCM. Meta-analysis revealed 28 loci (13 independent genome-wide significant variants (p-value $< 5\times10^{-8}$) and 16 <5% local false discovery rate variants (p-value $< 1.82\times10^{-6}$)). A genetic risk score (GRS) assessed the aggregate impact of these independent common variants: HCM risk was halved for individuals in the lowest quintile and more than doubled for those in the highest quintile.

Collectively, these analyses reject the null hypothesis that the genetic aetiology of HCM is restricted to known rare variant contributions and extend understanding regarding the genetic architecture of HCM.

The genetic architecture of hypertrophic cardiomyopathy

Dr. Andrew R. Harper

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

To Shalini and Reena

Acknowledgements

Many individuals and organisations have helped me towards my goal of better understanding the genetic architecture of hypertrophic cardiomyopathy.

I am thankful for the support and guidance Professor Martin Farrall and Professor Hugh Watkins have provided across multiple domains. Their expertise in statistical genetics and clinical medicine has helped refine my hypotheses and provided me with a supportive environment in which to grow as an independent researcher. The feedback provided by members of my Thesis Committee, particularly Professor Julian Knight, Professor Andrew Wilkie and Dr Anuj Goel, has helped drive forward my research efforts.

The University of Oxford, specifically the Wellcome Centre for Human Genetics and Radcliffe Department of Medicine, provided an unparalleled research environment that has been fundamental to both the research pursued, and my development as a researcher. There are several individuals from the University of Oxford I would like to thank, specifically: Dr Anuj Goel, Dr Kate Thomson, Dr Silvia Salatino, Dr Adam Waring, Dr Katja Gehmlich, Dr Liz Ormondroyd and Dr Chris Grace. Collectively these individuals helped me navigate the complexities of large scale bioinformatic and statistical analyses, better comprehend the molecular basis of hypertrophic cardiomyopathy and nuances of variant classification, appreciate the intricacies of clinical diagnostic testing, and consider the possible ethical implications of this research.

This research has benefited from strong collaborations with colleagues from numerous organisations including : Oxford University Hospitals NHS Foundation Trust Medical Genetics Laboratory (Michael Bowman), University of Oxford Wellcome Centre for Human Genetics (Mark McCarthy, Anubha Mahajan), the Hypertrophic Cardiomyopathy Registry (Chris Kramer and Stefan Neubauer), the NIHR BioResource for Rare Disease, Genomics England, UK Biobank, Royal Brompton Hospital (James Ware), Amsterdam University Medical Centre (Connie Bezzina) and Queen Mary University of London (Steffen Petersen). This research would not have been possible without the financial contributions provided by the Medical Research Council.

I am especially thankful for the constant support provided by my family, in particular my wife and parents.

Statement of authorship

The works submitted here for the degree of Doctor of Philosophy are my own, except where specifically stated within the text. This includes:

1. Dr Kate Thomson (Oxford Medical Genetics Laboratory): provided assistance with variant classification and in the development of a framework to dichotomise variants of uncertain significance for the purposes of genome-wide association analyses.

2. Dr Silvia Salatino (Wellcome Centre for Human Genetics): provided assistance with the conversion of fastq files (from gene-panel sequencing instruments) to BAM files and the required quality control.

3. Michael Bowman (Oxford Medical Genetics Laboratory): DNA extraction and preparation of gene-panel sequencing and genotyping of the HCMR cohort.

4. Dr Anuj Goel (Wellcome Centre for Human Genetics): performed pairwise GWAS analysis between sarcomere positive and sarcomere negative HCM.

Published works

Peer-reviewed publications

1. **Harper AR**, Bowman M, Hayesmoore JBG, Sage H, Salatino S, Blair E, Campbell C, Currie B, Goel A, McGuire K, Ormondroyd E, Sergeant K, Waring A, Woodley J, Kramer CM, Neubauer S, Farrall M, Watkins H, Thomson KL. **A Re-evaluation of the South Asian *MYBPC3*$^{\Delta 25}$ Intronic Deletion in Hypertrophic Cardiomyopathy.** *Circ Genom Precis Med.* 2020 Mar 12;. doi: 10.1161/CIRCGEN.119.002783. [Epub ahead of print] PubMed PMID: 32163302.

2. Neubauer S, Kolm P, Ho CY, Kwong RY, Desai MY, Dolman SF, Appelbaum E, Desvigne-Nickens P, DiMarco JP, Friedrich MG, Geller N, **Harper AR**, Jarolim P, Jerosch-Herold M, Kim DY, Maron MS, Schulz-Menger J, Piechnik SK, Thomson K, Zhang C, Watkins H, Weintraub WS, Kramer CM. **Distinct Subgroups in Hypertrophic Cardiomyopathy in the NHLBI HCM Registry.** *J Am Coll Cardiol.* 2019 Nov 12;74(19):2333-2345. doi: 10.1016/j.jacc.2019.08.1057. PubMed PMID: 31699273; PubMed Central PMCID: PMC6905038.

3. Thomson KL, Ormondroyd E, **Harper AR**, Dent T, McGuire K, Baksi J, Blair E, Brennan P, Buchan R, Bueser T, Campbell C, Carr-White G, Cook S, Daniels M, Deevi SVV, Goodship J, Hayesmoore JBG, Henderson A, Lamb T, Prasad S, Rayner-Matthews P, Robert L, Sneddon L, Stark H, Walsh R, Ware JS, Farrall M, Watkins HC. **Analysis of 51 proposed hypertrophic cardiomyopathy genes from genome sequencing data in sarcomere negative cases has negligible diagnostic yield.** *Genet Med.* 2019 Jul;21(7):1576-1584. doi: 10.1038/s41436-018-0375-z. Epub 2018 Dec 11. PubMed PMID: 30531895; PubMed Central PMCID: PMC6614037.

Pre-print publications

4. Waring AJ, **Harper AR**, Salatino S, Kramer CM, Neubauer S, Thomson KL, Watkins H, Farrall M. **Data-driven modelling of mutational hotspots and in-silico predictors in hypertrophic cardiomyopathy** *BioRxiv* doi: https://doi.org/10.1101/826164

International oral presentations

5. **Harper AR**, Thomson K, Mackley M, Watkins H, Ormondroyd E. **Secondary inherited cardiac condition findings from genome sequencing: Variant interpretation, assessment of phenotype and impacts of disclosure.** *American Society of Human Genetics* 2019 Houston, Texas, United States of America

Abstract

Hypertrophic cardiomyopathy (HCM) is the most common genetic heart disorder, affecting at least 1 in 500 individuals, and a leading cause of sudden death. Genetic testing for rare, causal, genetic variants in sarcomere genes is the standard of care and conducted at scale. However, more than half of HCM patients do not carry identifiable pathogenic variants and, in those that do, there is substantial variation in penetrance and disease expression.

Here, the genetic architecture of HCM is further evaluated, under a central hypothesis that the genetic aetiology of HCM extends beyond known rare variant contributions. Through a series of case-control analyses monogenic, oligogenic and polygenic models of disease were assessed. Burden testing analyses support prior knowledge regarding the monogenic basis to HCM. Quantitative analyses directed towards quantifying the penetrance and expressivity of disease-causing HCM variants were largely underpowered. Similarly, systematic evaluation for oligogenicity was underpowered. However, haplotype analysis of a candidate variant (*MYBPC3*Δ25) presumed to be of importance to oligogenicity revealed synthetic association with a rare pathogenic variant (*MYBPC3* c.1224-52G>A), quelling this specific oligogenic hypothesis. Polygenicity was evaluated through genome wide association analyses. The additive effects of common variants explained 34.0$\pm$2.4% of phenotypic variance in sarcomere-negative HCM, and 15.8$\pm$3.8% in sarcomere-positive HCM. Meta-analysis revealed 28 loci (13 independent genome-wide significant variants (p-value $< 5\times10^{-8}$) and 16 <5% local false discovery rate variants (p-value $< 1.82\times10^{-6}$)). A genetic risk score (GRS) assessed the aggregate impact of these independent common variants: HCM risk was halved for individuals in the lowest quintile and more than doubled for those in the highest quintile.

Collectively, these analyses reject the null hypothesis that the genetic aetiology of HCM is restricted to known rare variant contributions and extend understanding regarding the genetic architecture of HCM.

Contents

List of Figures

List of Tables

List of Abbreviations

ACMG . . . . American College of Medical Genetics and Genomics.

ACTC1 . . . . Actin, alpha cardiac muscle 1.

ACTN2 . . . . Actinin Alpha 2.

ADPRHL1 . . ADP-Ribosylhydrolase Like 1.

AFR . . . . . African

AHA . . . . . American Heart Assoication.

ALPK3 . . . . Alpha Kinase 3.

AMR . . . . . Ad Mixed American

ARVC . . . . . Arrhythmogenic right ventricular cardiomyopathy.

BAG3 . . . . . Bcl2-associated athanogene 3.

bp . . . . . Basepair

BRRD . . . . . BioResource for Rare Disease.

CI . . . . . Confidence interval.

CMR . . . . . Cardiac Magnetic Resonance imaging.

CSRP3 . . . . Cysteine And Glycine Rich Protein 3.

DNA . . . . . Deoxyribonucleic acid.

DCM . . . . . Dilated cardiomyopathy.

EAS . . . . . East Asian.

ECG . . . . . Electrocardiogram.

ESC . . . . . European Society of Cardiology.

EUR . . . . . European.

ExAC . . . . . Exome Aggregation Consortium.

FLNC . . . . . Filamin-C.

FHOD3 . . . . Formin Homology 2 Domain Containing 3.

GLA . . . . . Alpha-galactosidase.

gnomAD . . . genome Aggregation Database.

GP . . . . . General Practitioner.

GWAS . . . . . Genome wide association study.

HADS . . . . . Hospital Anxiety and Depression Scale.

HCM . . . . . Hypertrophic cardiomyopathy.

HCMR . . . . Hypertrophic Cardiomyopathy Registry.

HES . . . . . Hospital Episode Statistics.

HFpEF . . . . Heart failure with preserved ejection fraction.

HSPB7 . . . . Heat Shock Protein Family B (Small) Member 7.

HWE . . . . . Hardy-Weinberg Equilibrium.

ICC . . . . . Inhertied cardiac condition.

indel . . . . . insertion or deletion of bases in the genome.

JPH2 . . . . . Junctophilin 2.

kb . . . . . Kilobase.

LAMP2 . . . . Lysosomal associated membrane protein-2.

LGE . . . . . Late gadolinium enhancement.

LOD . . . . . Logarithm of Odds.

LP . . . . . Likely pathogenic.

LVH . . . . . Left ventricular hypertrophy.

LVNC . . . . . Left ventricular non-compaction cardiomyopathy.

MAF . . . . . Minor allele frequency.

MICRA . . . . Multidimensional Impact of Cancer Risk Assessment.

MIGRA . . . . Multidimensional Impact of Genomic Risk Assessment.

MYBPC3 . . Myosin binding protein C.

MYH7 . . . . β-myosin heavy chain.

MYL2 . . . . . Myosin regulatory light chain 2.

MYL3 . . . . . Myosin regulatory light chain 3.

NHS . . . . . National Health Service.

NIHR . . . . . National Institute for Health Research.

OMGL . . . . Oxford Medical Genetics Laboratory.

OR . . . . . Odds ratio

P . . . . . Pathogenic

PCA . . . . . Principal components analysis.

PLN . . . . . Phospholamban.

PRKAG2 . . Protein Kinase AMP-Activated Non-Catalytic Subunit Gamma 2.

SAS . . . . . South Asian.

SCARFE . . . Secondary Cardiac Findings Evaluation.

SLC6A6 . . . Solute carrier family 6, member 6

SNV . . . . . Single nucleotide variant.

T2DM . . . . . Type 2 diabetes.

TNNI3 . . . . Troponin I3.

TNNT2 . . . . Troponin T2, Cardiac Type.

TTN . . . . . Titin.

TTR . . . . . Transthyretin.

TPM1 . . . . . Tropomyosin 1.

UK . . . . . United Kingdom.

UKBB . . . . . United Kingdom BioBank.

UKAS . . . . . United Kingdom Accreditation Service.

VEP . . . . . Variant Effect Predictor.

vcf . . . . . Variant Call Format.

VUS . . . . . Variant of uncertain significance.

1 Introduction

Contents

1.1 Background

Hypertrophic cardiomyopathy (HCM) is the most commonly inherited cardiac condition (ICC). It is characterised by the presence of unexplained left ventricular hypertrophy (LVH), and molecularly defined as a disease of the cardiac sarcomere.[1, 2] Although defined as a rare disease, HCM is relatively common throughout the population. With a disease prevalence of $\sim$1/500, HCM appears relatively consistent

across major continents and is presumed to be equally distributed between men and women.[3–7] However, contemporary HCM case series have demonstrated a preponderance of men.[8]

HCM was traditionally recognised as a leading cause of sudden cardiac death in the young.[9] Although data suggest sudden cardiac death (mostly attributable to malignant ventricular arrhythmia) affects less than 1% of those diagnosed with HCM, this does not convey how heterogeneous an individual's propensity to develop sudden cardiac death can be [10, 11]. Consequently, in clinical practice the European Society of Cardiology's risk score is routinely adopted to partition individuals, based on seven criteria (age, family history of sudden cardiac death, unexplained syncope, left ventricular outflow gradient, maximum left ventricular wall thickness, left atrial diameter, and non-sustained ventricular tachycardia), into low (<4%), intermediate (≥4 to <6%) and high (≥6%) risk of sudden cardiac death to direct recommendations regarding the implantation of a cardiac defibrillator [12]. Although most affected individuals experience a normal life expectancy, there is a substantially increased risk of heart failure, thromboembolic disease and arrhythmia [13]. Nevertheless, as sudden cardiac death can be the first presentation of disease, clinical management strategies aimed at identifying and subsequently risk-stratifying 'at-risk' individuals have been established. Molecular genetics is central to this anticipatory approach and is performed at scale.[14] As almost all genetic variants causal for HCM demonstrate an autosomal dominant inheritance pattern, the clinical evaluation and subsequent management of affected individuals not only influences the proband, but also their extended family. However, despite a family possessing the same causal variant, the phenotypic features and clinical sequelae for each individual can vary considerably, reflecting the variable penetrance and expressivity of causal HCM variants. Moreover, for the majority of affected individuals, no causal variant is identifiable despite iterative improvements in DNA sequencing technology, and advances in variant classification practices.

Findings derived from studies investigating the genetic architecture of HCM can directly impact families affected by HCM, and also influence the decisions that

healthcare systems make regarding their care.[15] Better understanding of the genetic aetiology of HCM may reveal insights into critical biology of relevance not only to HCM, but potentially to the wider collection of myocardial phenotypes, including relatively common clinical presentations such as heart failure with preserved ejection fraction (HFpEF). HFpEF represents a heterogeneous group of conditions, unified by the presence of a supra-normal ejection fraction (i.e. left ventricular ejection fraction $\geq$50%), that are characterised by a disruption in diastolic functioning and incorporate phenotypic features beyond the myocardium.[16, 17] Such discoveries may one day stimulate the development of disease-modifying therapeutics that target underlying molecular processes, and transform the clinical management of patients from the palliation of symptoms and disease complications, to an era where disease is actively anticipated and prevented.

Beyond myocardial biology, issues pertaining to the genetics of HCM can provide perspective for many wider issues, ranging from the methods used to dissect the genetic aetiology of a rare disease, through to the ethical and societal challenges encountered within genomic medicine.

In this chapter, I aim to provide a background summary regarding the genetic architecture of HCM from both a clinical and scientific perspective. This will involve an initial clinical description of HCM, including a review of the known monogenic causes of HCM and clinical genetic approaches, before considering the methods that can be leveraged to further dissect the genetic basis of HCM.

1.2 Hypertrophic cardiomyopathy

Although anatomists of the 17th century described a phenotype resembling HCM, it was not until 1958 that Donald Teare, a forensic pathologist from the University of London, outlined the clinicopathological features of HCM in an 8-case autopsy series of young adults with asymmetric myocardial hypertrophy and sudden cardiac death.[18–21] Teare's report, now considered a landmark paper, unified many of the cardinal clinical features alongside histopathological hallmarks (myocyte hypertrophy, disarray and fibrosis) and was prescient in suggesting HCM as a

heritable condition[18]. In current times, HCM is recognised as a primary myocardial disorder, phenotypically characterised by the presence of myocardial hypertrophy, with hyperdynamic contraction, poor relaxation and increased energy consumption characterising its molecular profile.[22]

1.3 Clinical features

HCM can present in the context of extreme circumstances (aborted sudden cardiac death), or in more prosaic circumstances through cascade screening within an affected family, or increasingly via an incidental finding, detected through imaging or genetic testing. In the 60 years that have elapsed since Teare's original report, a series of technological advances in both cardiac imaging and molecular genetics have benefited the diagnosis and management of HCM, contributing towards a reduction in its annual mortality from 3-6% in the 1980s to ∼0.5% in the 2000s.[10, 23–25]

Symptoms and clinical examination alone confer limited sensitivity and specificity in establishing a diagnosis of HCM. Symptoms may include dyspnoea, fatigue, chest pain, palpitations and pre-syncopal/syncopal episodes, but over 80% of individuals diagnosed with HCM are asymptomatic. Pathophysiological mechanisms responsible for these symptoms include impaired cardiac filling/diastolic function and/or emptying, microvascular dysfunction and arrhythmia.[12] Similarly, physical examination may reveal a bisferiens pulse, a forceful and sustained apical impulse, an audible fourth heart sound and a crescendo-decrescendo systolic murmur at the lower left sternal edge (that intensifies with the Valsalva manoeuvre). However, these clinical signs tend to reflect the severity of left ventricular outflow tract obstruction which may be absent in a third of affected individuals. A range of electrocardiographical features may be noted, including atrial fibrillation, pathological Q waves, left atrial abnormalities, LVH by voltage criteria and/or widespread ST-T wave changes.[12] Several machine learning approaches have been attempted to either diagnose, or partition HCM into discrete subtypes based on ECG features.[26–28]

Central to any HCM diagnosis is the presence of unexplained LVH in the absence of loading conditions, infiltrative or storage disorders, and/or haemodynamic stress

[12, 29]. Consensus regarding the specific wall thickness measurement that defines LVH in HCM remains elusive; the European Society of Cardiology (ESC) opt for ≥15mm, or ≥13mm in the context of a positive family history, whereas the American Heart Association (AHA) apply a generic threshold of ≥13mm.[12, 29] In ∼70% of HCM cases, hypertrophy is asymmetric and localised to the basal anterior septum and contiguous left ventricular free wall.[30]

Whilst wall thickness can be measured using echocardiography, the diverse phenotypic features associated with HCM are better captured by cardiac magnetic resonance (CMR) imaging. CMR can characterise histopathological hallmarks of HCM, including hypertrophy, both the extent and morphological distribution; fibrosis, through late gadolinium enhancement (LGE) and extracellular volume mapping; and most recently, myocyte disarray, through use of diffusion tensor CMR and fractional anisotropy.[30–33] Advances in CMR have facilitated improvements in diagnostic accuracy and risk stratification of HCM.[34]

1.4 HCM as a heritable condition

The realisation that HCM was a heritable condition, demonstrating autosomal dominant inheritance patterns in ∼60% of affected individuals, facilitated the identification of the first genes ((*ACTC1* (Actin, alpha cardiac muscle 1, detected in 1999), *MYBPC3* (myosin binding protein C, detected in 1990), *MYH7* (*β*-myosin heavy chain, detected in 1989), *MYL2* (myosin regulatory light chain 2, detected in 1996), *MYL3* (myosin regulatory light chain 3, detected in 1996), *TNNT2* (Troponin T2, Cardiac Type, detected in 1994), *TNNI3* (Troponin I3, detected in 1997) and *TPM1*(Tropomyosin 1, detected in 1994)) associated with HCM during the late 1980s and 1990s using family based linkage analysis.[35–41] Family based linkage analysis is a methodology that identifies loci implicated in disease through the co-segregation of microsatellite markers with affected status across multi-generational family pedigrees. Collectively these findings established dogma that HCM was a monogenic, autosomal dominant condition of the cardiac

sarcomere and tempered competing theories of the time that HCM was a collection of aetiologically distinct diseases.

Sarcomeres are the most basic contractile units of striated muscle. The sarcomere is composed of interdigitating thick (myosin molecules) and thin (actin molecules) filaments, and through the interaction of these molecular subunits, adenosine triphosphate is hydrolysed and mechanical force generated, as outlined by Huxley's sliding filament theory.[42, 43]

In modern day practice DNA sequencing has become a cornerstone to the delivery of care for patients with HCM. One use has been in the identification of patients who present with concentric LVH attributable to a HCM phenocopy gene (rather than sarcomeric HCM). In this context, the term phenocopy is used to describe a condition that appears phenotypically similar to that of HCM, but is underpinned by a different molecular mechanism and therefore managed differently. One such example is Fabry disease, attributable to disease-causing variants in alpha-galactosidase (*GLA*), where treatment comprising of enzymatic replacement therapy is available.[44] HCM phenocopies account for ∼2% of patients referred for HCM genetic testing, and alongside Fabry disease, conditions include rare inherited metabolic disorders such as Danon disease (lysosomal associated membrane protein-2, *LAMP2*), storage disorders like *PRKAG2*(Protein Kinase AMP-Activated Non-Catalytic Subunit Gamma 2)-related cardiomyopathy, infiltrative disorders such as hereditary amyloidosis (transthyretin, *TTR*), primary mitochondrial diseases, neuromuscular disorders and other genetic syndromic disorders.[45–48]

In most cases, DNA sequencing provides an opportunity to identify causal genetic variants and facilitate family-based predictive genetic testing, so as to direct care towards family members at genetic risk of developing HCM. The vast majority of disease-causing variants are detected in cardiac sarcomere genes, with over >85% of genetic causes attributed to either presumed loss-of-function variants, most notably truncating variants, in *MYBPC3*, or missense variants in *MYH7*.[8] Carriers of disease-causing variants in *MYH7* typically demonstrate high disease penetrance and moderate to severe LVH, whereas *MYBPC3* carriers tend to demonstrate a less

severe phenotype that emerges in mid-to-late adulthood.[1] Disease mechanisms underpinning HCM attributable to disease-causing variants in *MYH7* and *MYBPC3* have been reported through biophysical experiments evaluating how disease causing variants disrupt myosin conformations in the sarcomere.[49, 50]

Having been introduced over 15 years ago, genetic testing is now recommended by both the ESC and AHA guidelines.[12, 29] Consequently, large national repositories of clinically referred patient samples have been established, and in combination with large reference datasets (e.g. ExaC, gnomAD or TOPMed) and application of contemporary variant classification methodologies, insights into the genetic aetiology of HCM have been gleaned.[8, 51–54] This includes the realisation that only a minority of individuals clinically diagnosed with HCM harboured likely pathogenic or pathogenic variants in well established HCM genes.[8, 55] Furthermore, many genes previously included on clinical diagnostic testing panels appeared to show no difference in the rate of rare genetic variation between cases and controls. Second, whilst the advancement of variant classification has been essential in reducing the number of false positives communicated to patients, the framework developed by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) may be overly conservative in assigning pathogenic status. This observation stems from the realisation that individuals yielding a variant of uncertain significance tend to follow a more-similar trajectory, in terms of clinical outcomes, as those with likely pathogenic/pathogenic variants.[56] Additionally, data derived from two large clinical genetics services indicates two-thirds of variants of uncertain significance are in fact causal.[8]

1.5 Clinical genetic screening

The clinical genetic testing strategy employed to identify disease-causing variants associated with HCM has traditionally been contingent on the clinical rationale for testing. When a disease-causing variant was identified within a family, sequencing of that specific variant was performed across first-degree relatives, using Sanger's method, a process known as cascade screening.[14, 57] However, this procedure

is reliant on first identifying a disease-causing variant. To achieve this, probands would traditionally undergo gene-panel sequencing, a process that involves the selective capture and amplification of specific exons and untranslated regions across a collection of well-defined genes, causal for either sarcomeric HCM or an HCM phenocopy, prior to DNA sequencing.

1.5.1 HCM associated genes

Several working groups have been assembled in an attempt to provide consensus regarding the genes that should be included on clinical diagnostic gene panels and prioritised for variant interpretation.[58, 59] Which genes to test on an HCM gene-panel remains undecided. Eight core sarcomere genes (*ACTC1*, *MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *TNNT2*, *TNNI3* and *TPM1*), initially identified through linkage analysis in multi-generational families and subsequently confirmed through large-scale burden testing, account for almost all cases yielding a disease-causing variant.[8, 35–41] Seven additional genes (*ACTN2* (Actinin Alpha 2), *ALPK3* (Alpha Kinase 3), *CSRP3* (Cysteine And Glycine Rich Protein 3), *FHOD3* (Formin Homology 2 Domain Containing 3), *FLNC* (Filamin-C), *JPH2* (Junctophilin 2) and *PLN* (Phospholamban)) have also been deemed to represent genes capable of harbouring disease-causing variants, but in aggregate account for less than 1% of cases and are often challenging to interpret given the poor signal:noise ratio.[60–65] HCM phenocopy genes (*GLA*, *PRKAG2*, *LAMP2* and *TTR*) tend to be included to exclude common differential diagnoses.[45–48]

1.5.2 DNA sequencing

DNA sequencing is a methodology that aims to determine the order of four nucleotide bases along a stretch of DNA. DNA sequencing methods have evolved and broadly, three generations of DNA sequencing technology exist. The first generation of DNA sequencing technologies relate to semi-automated implementations of Sanger's method.[57] Sequencing technologies that are referred to as high-throughput or massively parallel tend to represent the second generation of sequencing platforms,

with Illumina's sequencing-by-synthesis methodology recognised as the predominant approach. Unless otherwise stated, second generation sequencing approaches were used throughout this thesis. Technical details regarding the background and methodological considerations for second generation sequencing are provided elsewhere.[66–69] In brief, second generation sequencing, specific to the Illumina protocol, requires three steps to be performed: 1) library preparation; 2) sequencing using reversible terminator technology; and 3) bioinformatic manipulation. Library preparation requires extracted DNA to be first fragmented and then enzymatically modified to allow the single-stranded DNA fragments to adhere to a sheet of glass (known as the flowcell) within the sequencing platform. Once adhered, each DNA fragment is copied multiple times, in a process known as amplification, for the creation of clonal clusters. The specific nucleotide sequence for each DNA fragment can then be reported as fluorescently-labelled nucleotide bases become incorporated with the complementary DNA fragments.[70] This approach generates relatively short reads (i.e. 50 to 300 nucleotides in length), that are subsequently aligned to a reference genome and then evaluated for variants (further discussion below).

Most high-throughput germline sequencing technologies implement short read approaches. Third generation sequencing approaches, that encompass platforms pursuing real-time, single-molecule approaches, can generate longer-reads (discussed below). Most clinical diagnostic genetic testing approaches utilise second generation sequencing platforms, either with gene panel, exome or genome sequencing. Differences between gene panel, exome and genome sequencing approaches influence the practical use of each technology.

Traditionally gene-panel sequencing, limited to robustly associated HCM genes, would be deployed as a first line test, and only in situations where HCM was part of a more complex phenotype, or an atypical presentation (i.e. early age of onset, or mixed cardiomyopathy phenotypes within a family) would an alternative sequencing strategy be considered. This would have typically involved consideration for a wider gene-panel or a genome-wide sequencing approach, either capturing the entire genome or limited to the 2% portion that encodes for proteins and

non-coding RNA (i.e. the exome). However, with dramatic reductions in sequencing costs, clinical diagnostic testing is increasingly undertaking genome-wide sequencing approaches as a first line diagnostic test, with genomic regions not pertinent to the clinical question initially masked.[71] Such an approach aims to optimise the signal-to-noise ratio, given the vast abundance of rare variation that is present across the human genome: the genome consists of ∼3.2 billion base pairs, and the average European individual harbours ∼30,000 rare (MAF $\leq$ 0.01) variants, of which twenty are protein-truncating.[53, 72] By only selecting genomic regions of relevance to the disease of interest, either using a gene panel or a masking approach, rare variants in genomic regions, robustly associated with the disease of interest, can be prioritised for variant interpretation.

When masking approaches are adopted exome sequencing tends to be preferentially selected. Exome sequencing captures protein-coding regions enriched for disease-causing variants (>85% occur in, or close to, protein-coding regions).[73] It is acknowledged that non-coding genomic regions, neglected by exome sequencing capture techniques, harbour additional disease-causing variants that are detectable via genome sequencing.[74–77] For example, deep intronic regions of *MYBPC3* have been shown to harbour disease-causing variants in HCM that would have been otherwise missed by exome sequencing.[74] Beyond HCM, genome sequencing has been able to identify novel causes of disease. For instance, *de novo* variants in regulatory elements have been shown to cause neurodevelopmental disorders, and a frameshift variant (-65-66insT) in the 5 prime untranslated region of *NF2* that disrupts upstream open reading frames.[76, 77] Exome sequencing is also substantially cheaper than genome sequencing. Cost is particularly important when family-based sequencing approaches are pursued; exome sequencing can be performed for several family members for the equivalent cost of a single individual undertaking genome sequencing.

There are additional differences between exome and genome sequencing approaches. Exome sequencing tends to generate higher coverage (i.e. the number of times a base is correctly covered by an amplified read) than genome sequencing,

albeit with greater variability. High coverage improves the probability that the underlying base is called correctly, as measured by the 'Phred' score. Bases demonstrating a probability of error >1%, equivalent to a Phred score <Q20, are typically discarded. Although high coverage does not automatically confer high certainty, particularly if non-independent duplicate reads are present. Whilst exome sequencing may confer higher coverage, genome sequencing offers greater sensitivity for the detection of structural variants (i.e. insertions, deletions, duplications, inversions, or translocations $\geq$ 50 bp in length) and avoids potential bias incurred by target capture steps, such as reference bias. Reference bias occurs due to target capture assays being optimised for bases present in the reference sequence and not risk-associated alleles.

It is acknowledged that long read (i.e. reads with a median length of 5-10 kb) DNA sequencing could theoretically improve several aspects of clinical diagnostic genetic testing, but at present, remains prohibitively expensive. Long read sequencing has the potential to improve the characterisation of structural variants, paralogous regions or highly repetitive sequences. In addition, long read sequencing has the ability to directly phase variants several kilobases apart, particularly in the setting of compound heterozygosity, where two variants are co-located in the same gene.[14, 78–83]

1.5.3 Reference genome and variant calling

Before variants can be identified, the large volume of reads and accompanying quality scores generated by the sequencing instrument need to be processed. This can either involve a *de novo* assembly approach, or more likely, an alignment process against a reference genome based on the Burrows-Wheeler Transform.[84] Whilst GRCh37 is frequently used as the reference genome, it is haploid in nature, contains many risk-associated alleles and has a disjointed haplotype structure, having been derived from thirteen anonymous volunteers.[78, 85] The latest reference genome produced by the Genome Reference Consortium, GRCh38, overcomes some of these challenges but has not been universally adopted since its release in 2013.

For the purposes of benchmarking sequencing pipelines and variant calling tools, the National Institute of Standards and Technology, specifically the Genome in a Bottle consortia, have developed a set of reference resources including a set of high-confidence genotype calls.[86–88]

Variants are identified when differences between the reference genome and aligned reads emerge. Numerous bioinformatic software tools have been developed to try and correctly identify short germline variants from sequence data, including the Genome Analysis Toolkit (GATK).[89] It is acknowledged that variable accuracy exists in correctly identifying types of variants: most software tools can accurately call SNVs, but struggle to offer a similar level of accuracy for indels.[14] Differences that are detected between the sequenced reads and reference genome are stored in a variant call format (VCF) file.

1.5.4 Variant classification

Prior to the release of the ACMG guidelines in 2015, many variants were found to have been assigned causal status without sufficient supporting evidence, and clinical decisions based on such variants had the potential to harm patients (i.e. inappropriate diagnosis, discharge or intervention). Reports emerged of these misclassifications through ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), a publicly accessible database that aims to capture genotype-phenotype relationships. A key contributing factor to these misclassification can be attributed to naïve assumptions regarding the abundance of rare variation across the genome. Until the emergence of large-scale reference datasets such as ExAC or gnomAD, variants detected in cases were often considered causal if they did not appear in a control cohort of several hundred anonymous blood donors.[52, 53] Analysis performed in ExAC, a collection of ∼60,000 human exomes, demonstrates how prevalent coding variation is, with one in every eight coding basepairs, on average, being a source of variation. [53, 90]

The introduction of the ACMG/AMP guidelines helped to remedy this situation by providing a systematic framework, consisting of 28 rules, that evaluates variants

for evidence of pathogenicity across multiple domains, before assigning one of five different classifications, from benign to pathogenic.[91] As the ACMG/AMP guideline was developed to facilitate variant interpretation across all Mendelian disease genes, guidance provided by the framework is often generic, relatively subjective in parts, and susceptible to misinterpretation, as evidence by the presence of substantial heterogeneity between independent variant classification assessors.[92]

In an attempt to resolve ambiguous guidance and capture the intricacies associated with specific genes and disease areas, disease-specific specifications have emerged from the ACMG/AMP guidelines. For example, the cardiomyopathy expert panel, convened by Clinical Genome Resource (ClinGen), has published an adaptation of ACMG/AMP classification framework specific to *MYH7*.[93] Of the original 28 ACMG/AMP framework rules, 9 were considered not to be applicable and 12 required disease and/or gene specific adjustments. This includes the re-calibration of allele frequency thresholds, above which variants are determined to be benign (rule BA1). The general ACMG/AMP guidelines suggests a variant with an allele frequency in excess of 5% is benign, but in the context of HCM, where disease prevalence is 1/500, this threshold is too lenient. Instead, an allele frequency threshold of 0.1% is more appropriate, but this is still more lenient than what most research studies investigating HCM have typically adopted (i.e. a threshold of 0.01%). Nevertheless, use of 0.1% as an allele frequency threshold is based on the allele frequency of the most frequently detected pathogenic variant in HCM (*MYBPC3* p.R502W) in a reference databases, such as gnomAD.[8, 94]

Furthermore, the 2015 ACMG/AMP guideline is acknowledged to be overly conservative, with a large proportion of variants deemed 'uncertain significance'. This can occur when specific ACMG/AMP rules are conflicting. To overcome this challenge the next iteration of the ACMG/AMP guidelines is seeking to utilise a quantitative Bayesian framework to better synthesise orthogonal layers of evidence.[95, 96] Variant classification needs to therefore be viewed as a dynamic process, subject to change, particularly for variants currently assigned uncertain significance status. Whilst this may be understandable from a scientific perspective,

the constant threat of variant re-classification has the potential to invoke clinical repercussions. Furthermore, the clinical pathways that address the re-contacting of individuals who yield a variant that undergoes re-classification remains relatively embryonic and improved governance structures are necessary to safeguard patients from this avoidable harm.

1.6 Genetic architecture of hypertrophic cardiomyopathy

The first sarcomere genes harbouring pathogenic variants causal for HCM were identified in the 1990s. Since this time there have been continued efforts to further characterise the genetic architecture of HCM, largely through studies focusing on reductionist hypotheses, with an assumption that additional disease-associated genes exist as a cause for HCM in the sarcomere-negative population. However, it is also plausible that sarcomere-negative HCM is attributable to alternative genetic architectures, such as oligogenicity or polygenicity (Figure 1.1). In support of this is evidence that sarcomere-negative HCM confers a better prognosis and lower likelihood of family recurrence. Whilst both an oligogenic and polygenic model of HCM have been postulated, they remain relatively unexplored areas.

1.6.1 Monogenic

It is plausible that sarcomere-negative HCM is explained by rare variants in candidate genes that have yet to be discovered. A monogenic hypothesis assumes that sarcomere-negative HCM is genetically highly heterogenous, with rare causal variants nested along a long tail of genes implicated in disease. To investigate this two approaches have traditionally been employed. The first approach has used traditional family-based approaches, such as linkage analysis. This is challenging as most sarcomere-negative disease is confined to nuclear families or singleton cases, and it is extremely rare for large extended families to lack a genetic diagnosis. This observation alone suggests additional monogenic causes of HCM are unlikely. However, rather than using family-based discovery methods, further enquiry into

the monogenic basis of HCM has utilised gene-level burden testing, given the availability of DNA sequence data from HCM cases and controls. Gene-level burden testing refers to a statistical approach that assesses for evidence of enrichment or depletion, between cases and controls, for the aggregate total of qualifying variants (i.e. variants from a specific category, such as rare protein-truncating variants) from a predefined genomic region representative of a gene.[97] Burden testing approaches have been adopted, and directed towards genes with a high pre-test probability of disease involvement, given the multiplicity-adjusted significance threshold of 2.6×10^{-6} (i.e. $\alpha = 0.05/19{,}000$ protein-coding genes) required from exome-based, hypothesis free approaches. This approach has also been necessary, as whilst several thousand HCM cases have undergone gene-panel sequencing across a selection of cardiomyopathy-associated genes, relatively few HCM cases have available exome or genome sequencing data.

To date, seven genes (*ACTN2, ALPK3, CSRP3, FHOD3, FLNC, JPH2*, and *PLN*) have emerged in support of a monogenic hypothesis, beyond the core sarcomere genes, with supporting evidence derived either from variant co-segregation or robust case-control analyses.[60–65] However, the cumulative contribution these genes make towards HCM is negligible; a result that resonates with clinical observations. Efforts to further define the monogenic basis of sarcomere-negative HCM are likely to require extremely large case-control studies, upon which hypothesis-free burden testing can be performed, and/or family-based sequencing across unrelated families.

1.6.2 Oligogenic

An alternative hypothesis would be that sarcomere-negative HCM is the result of several variants, from across the allele frequency spectrum, each insufficient to cause disease in isolation, but by synergistically combining, critical biological pathways could be perturbed and lead to disease. This genetic model is considered to represent oligogenicity, and is reliant on the co-inheritance of multiple variants from both parents. However, despite numerous literature reports describing the co-occurrence of multiple variants in individuals with HCM, proving there is an

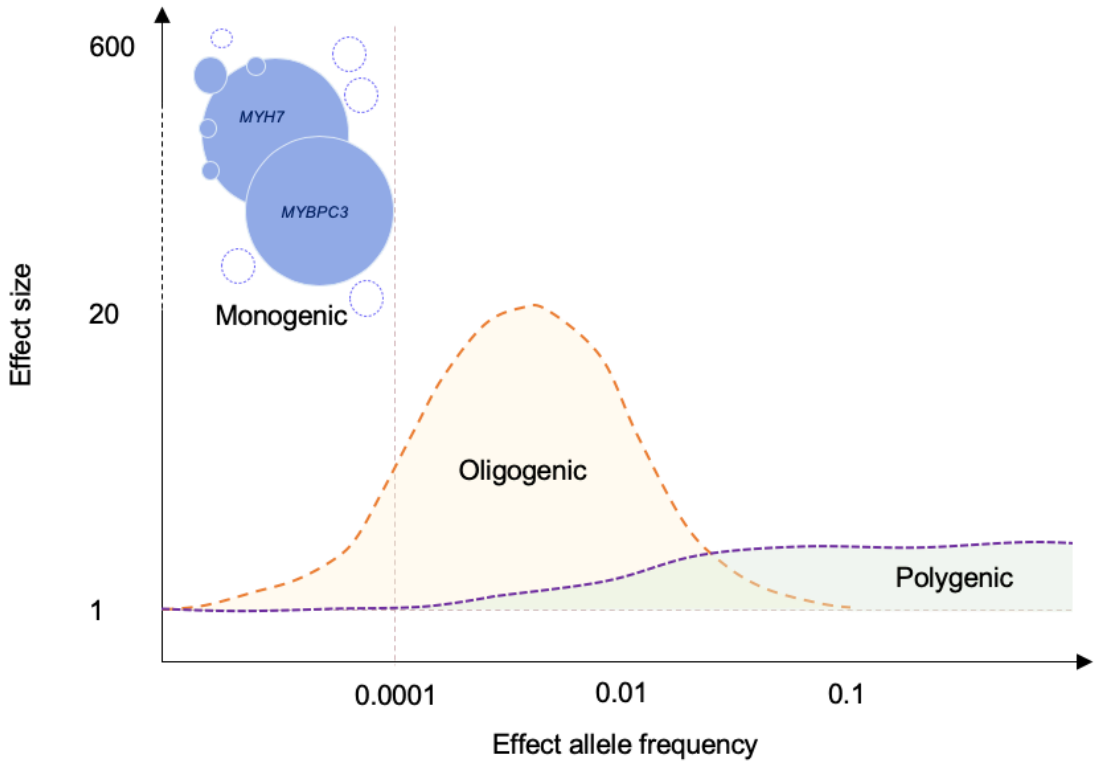


Figure 1.1: A schematic demonstrating the proposed relationship between allele frequency and effect size with respect to hypertrophic cardiomyopathy. Genes (i.e. *MYH7* and *MYBPC3*) supporting a monogenic model of hypertrophic cardiomyopathy (HCM) have been previously identified (Blue). Monogenic genes harbour pathogenic variants with high effect sizes and of low frequency within the general population (effect allele frequency less than 0.0001). Oligogenic genes have yet to be reported (yellow), but are presumed to span the allele frequency spectrum, with each oligogenic variant conferring a moderate effect size. Multiple common variants (typically an allele frequency > 0.01, but theoretically possible to be supported by rarer variants) of modest (i.e. odds ratio < 2) effect size are anticipated.

oligogenic basis to HCM has proven challenging. Most reports tend to present case-series data without consideration for the null distribution of variants in an unaffected, ancestrally matched, control cohort.

Evidence supporting an oligogenic basis to cardiomyopathy has been forthcoming from family-based exome sequencing studies in multiply affected nuclear families, specifically where low-frequency protein-altering variants are prioritised and characterised *in vivo*.[98] There is evidence oligogenicity contributes towards other rare diseases[99–105], and that the variants involved in disease need not be confined to the low-frequency or rare allele frequency spectrum. For example, common variants

in *cis* regulatory elements contribute towards an oligogenic model of disease in Hirschprung's disease, a rare developmental disorder of the enteric nervous system that affects 15/100,000 live births.[106] Examples of common variants (i.e. minor allele frequency > 1%) that contribute towards HCM are scarce, and when they do exist, tend to be confined to established cardiomyopathy genes: the most notable being a 25 base pair intronic deletion in *MYBPC3*, present in 4-8% of South Asian individuals, that confers a 7-fold increased risk of disease.[107]

Exploring possible digenic architectures, using a systematic gene-burden approach, is challenging. To test all possible pairwise gene combinations results in over 1.8×10^7 tests (based on the binomial coefficient $\binom{19{,}000}{2}$), and a multiplicity-adjusted significance threshold of 2.8×10^{-10} (i.e. $\alpha = 0.05/1.8\times10^7$ pairwise gene combinations). Given the prevalence of HCM, establishing a suitably sized cohort that facilitates this experimental design may not be possible in the near term, although this is dependent on how strong the hypothesised synergism is. Alternative strategies will need to be devised, such as the prioritisation of genes highly expressed in myocardial tissue.[108]

1.6.3 Polygenic

Since the first genome-wide association studies (GWAS) were performed in 2005 over 60,000 genome-wide significant association signals have been reported across a vast array of traits and diseases.[109] Findings from GWAS confirmed that susceptibility towards common complex diseases, such as type 2 diabetes or multiple sclerosis, is through the additive effects of many common genetic variants that are shared across the population (i.e. those with a minor allele frequency >1%), each of relatively small effect. Furthermore, it is likely that the aggregate additive effect of common variants also contributes towards disease susceptibility in rare disease. This is perhaps best exemplified by familial hypercholesterolaemia (FH), an autosomal dominant lipid disorder that affects ∼1 in 200 individuals and is characterised by elevated low-density lipoprotein cholesterol (LDL-c) levels and premature coronary artery disease.[110] For ∼40% of individuals with possible FH, as defined by the

Dutch Lipid Clinic Criteria, a disease-causing variant can be detected in one of three genes: low-density lipoprotein receptor (*LDLR*), apolipoprotein B-100 (*APOB*), and proprotein convertase subtulisin/kexin type 9 (*PCSK9*).[111, 112] By aggregating the contribution common genetic variants make towards elevated LDL-c, polygenic risk scores suggest up to 8% of the general population have at least a 3-fold increased risk for atherosclerotic cardiovascular disease.[113] Consequently, investigators have attempted to equate risk from both monogenic and polygenic causes as being equivalent.[113] Using data from the UK Biobank, Trinder et al. (2020) demonstrated that although both monogenic (n=277, hazard ratio (HR) = 1.93 [95% CI: 1.34-2.77]; p-value < 0.001) and polygenic causes (i.e. $>95^{\text{th}}$ percentile based on 223 single nucleotide polymorphisms) (n=2379; HR = 1.26 [95% CI: 1.03-1.55]; p-value = 0.03) of FH were associated with an increased risk of cardiovascular events, when compared against individuals with hypercholesterolaemia without a genetic cause, monogenic causes conferred the greatest risk. [114]

However, the clinical utility and role of polygenic risk scores remains unclear. Polygenic risk scores that advocate screening for individuals at a 3-fold increased risk of disease, such as those proposed by Khera et al (2018) for the identification of individuals at risk of coronary artery disease, may not be appropriate for all diseases, or even for the detection of coronary artery disease. For example, Nicholas J. Wald & Robert Old outline how the model proposed by Khera et al (2018) would confer an ∼85% false negative rate and 5% false positive rate.[113, 115] Application of this philosophy to other diseases, such as HCM, may not be appropriate and could lead to iatrogenic harm.

Polygenic risk scores may have greater utility in providing insights into disease biology; use of partitioned polygenic risk scores, in combination with Mendelian randomisation, may help dissect the aetiological pathways that underpin the intermediary phenotypes constituting a given disease.[116, 117]

Pursuit of a common variant hypothesis represents a departure from the classical rare variant, reductionist model that has dominated research investigating the genetic architecture of HCM. This theory has been partially tested in a previous

GWAS, although the experimental design was relatively underpowered, having only involved 174 HCM cases and 823 controls. Re-evaluation of this hypothesis is therefore required, particularly as the study suggested a common, intronic variant in *FHOD3*, rs516514, yields a relatively large effect (odds ratio (OR) = 2.45 [95% CI: 1.76–3.41]; p-value $< 1.25\times10^{-7}$) towards HCM risk.[118] Additionally, it could be hypothesised that common variants have differing roles in sarcomere-positive and sarcomere-negative disease. Whereas the aggregate burden of common genetic variants might explain why individuals develop sarcomere-negative HCM, in sarcomere-positive HCM, where disease risk is determined by the presence of a disease-causing variant, common variants could act to modify the phenotype, thus influencing the observed penetrance and expressivity.

1.7 Hypothesis and key objectives

Central to this thesis is the hypothesis that the genetic architecture of HCM extends beyond the rare variant contributions that have been extensively investigated to date. To assess this hypothesis four core objectives will be addressed. Outlined below are a list of core objectives, specific to each Chapter. Alongside the overall core objective, each Chapter aims to address a list of specific research questions, summarised here, below each core objective. A more detailed rationale for addressing these specific research questions is outlined in the introductory section of each Chapter. The core objectives for this thesis include:

- Evaluate HCM from a monogenic disease perspective (Chapter 3)

  Before I can evaluate whether the genetic architecture of HCM extends beyond a monogenic model of HCM, it is essential to appreciate the contribution rare variants make towards the null hypothesis (i.e. that the genetic architecture of HCM does not extend beyond a monogenic model of disease) in the HCM cohorts used throughout this thesis. As such, cohorts will be assessed in relation to: 1) the overall proportion of HCM cases that harbour disease-causing variants; 2) the proportion of individuals in whom multiple disease

causing variants are reported; 3) what the most frequently occurring disease-causing variants are; and 4) whether genes presumed to be involved in HCM show evidence of enrichment for rare, potentially disease-causing variants in HCM cases, compared with controls.

- Consider possible oligogenic causes of HCM (Chapter 4)

  Following the characterisation of cohorts using a monogenic model of disease, alternative genetic architectures will be considered. Chapter 4 considers whether HCM could be attributable to the co-inheritance of multiple variants, from across core sarcomere genes, and whether this supports an oligogenic model of disease.

- Assess the penetrance associated with disease causing variants in HCM (Chapter 5)

  Chapter 5 aims to answer several research questions that centre around the concept of penetrance in HCM. This includes: 1) an assessment for whether the UK BioBank is a suitable cohort to derive HCM penetrance estimates; 2) consideration for whether an approach that leverages large case-control cohorts can generate penetrance estimates that are be clinically informative; and 3) an assessment regarding the prevalence, and associated expressivity, specific to variants, known to cause HCM, that are detected in individuals without a personal or family history of HCM.

- Investigate the role of common variants in HCM (Chapter 6)

  Chapter 6 aims to explore whether there is evidence to support a polygenic model of HCM. As such, Chapter 6 evaluates: 1) the contribution common variants make towards HCM through heritability and genome-wide association analyses; 2) whether the contribution common variants make towards HCM is dependent on carriage of a disease-causing rare variant; and 3) should there be evidence of polygenicity, does the aggregate burden of these common variants increase an individual's likelihood of developing HCM.

2
Materials and Methods

Contents

2.1 Introduction

This chapter provides details regarding the materials and methods used throughout this thesis. This includes information relating to the case and control cohorts used, the reference data used, how DNA sequencing and genome-wide genotyping were performed, in addition to details relating to the bioinformatic pipeline and quality control procedures.

2.2 Cohorts

2.2.1 Overview

A table summarising the data sets compiled and used throughout this thesis is presented in Appendix A.1.

2.2.2 Oxford Medical Genetics Laboratory

The Oxford University Hospitals NHS Foundation Trust's Medical Genetics Laboratory (OMGL) is a large regional clinical genetic testing service that performs diagnostic testing for Mendelian disease, including inherited cardiac conditions (ICC). The OMGL cohort contains consecutive probands (i.e. the first member of a family that presents for the genetic evaluation of a disease), clinically diagnosed with an ICC referred for clinical genetic testing, between 2013 and 2018, by an ICC specialist.

Data generated by the OMGL clinical-diagnostic high-throughput sequencing pipeline was repurposed for research, adhering to data governance legislation enacted by Oxford University Hospitals NHS Foundation Trust.

Available next-generation sequence data for all ICC phenotypes, including 3,976 cardiomyopathy probands (HCM (n=2,757), dilated cardiomyopathy (DCM) (n=1,071), arrhythmogenic right ventricular cardiomyopathy (ARVC) (n=267), and left ventricular non compaction (LVNC) (n=52)) and arrhythmia probands (n=928), were retrieved and processed. For cardiomyopathy probands, the OMGL clinical diagnostic pipeline performed clinical grade gene panel sequencing, which initially comprised 27 cardiomyopathy associated genes (*ACTC1, ACTN2, ANKRD1* (Ankyrin Repeat Domain 1), *CRYAB* (Crystallin Alpha B), *CSRP3, DES* (Desmin), *DSC2* (Desmocollin 2), *DSG2* (Desmoglein 2), *DSP* (Desmoplakin), *FHL1* (Four And A Half LIM Domains 1), *FHL2* (Four And A Half LIM Domains 2), *GLA, LAMP2, LMNA* (Lamin A/C), *MYBPC3, MYH7, MYL2, MYL3, PKP2* (plakophilin-2), *PLN, PRKAG2, SCN5A* (Sodium Voltage-Gated Channel Alpha Subunit 5), *TMEM43* (transmembrane protein 43), *TNNI3, TNNT2, TPM1* and *TTN*(titin)). An additional eight genes were added to the panel in 2015 (*BAG3* (Bcl (B-cell lymphoma)-2-Associated Athanogene 3), *DMD* (dystrophin), *FLNC, RBM20* (RNA Binding Motif Protein 20), *TAZ* (Tafazzin), *TNNC1* (Troponin C1, Slow Skeletal And Cardiac Type), *TTR*, and *VCL* (Vinculin)). Although many of the genes included on the panel contribute towards more than one ICC, evaluating all 35 genes for each ICC attenuates the diagnostic specificity of the genetic test, and

consequently genes not robustly associated with a given ICC are masked when clinical-grade variant interpretation is performed.

For the purposes of this thesis, all 35 genes were considered for probands diagnosed with HCM. No phenotypic information, beyond a brief description of the diagnosis included on the referral, was available. Self-identified ancestry data were not available for this cohort.

2.2.3 Hypertrophic Cardiomyopathy Registry

The Hypertrophic Cardiomyopathy Registry (HCMR) cohort represents a prospective, multicentre, longitudinal, observational registry that aims to improve risk prediction for important adverse clinical outcomes in HCM (ClinicalTrials.gov identifier: NCT01915615; REC reference: 14/SC/0190).[119] The HCMR cohort recruited 2,762 individuals, aged 18 to 65 years old with evidence of unexplained left ventricular hypertrophy (wall thickness >15mm) from across 44 centres; 1,362 individuals were enrolled in North America and 1,400 in Europe. Demographic details, cardiac phenotyping (cardiovascular magnetic resonance imaging and echocardiography) and biomarker data were all collected at baseline and managed by an external clinical research organisation (MedStar Health Research Institute, Washington, USA). The genetic component was coordinated by the University of Oxford and includes both 35 gene panel sequencing and genome-wide genotyping. As genome-wide genotyping had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant.

2.2.4 BioResource for Rare Disease

The BioResource for Rare Disease (BRRD) cohort was established with the aim of discovering DNA sequence variants using next generation sequencing technologies to explain unresolved, rare diseases.[120] The BRRD was a precursor to the 100,000 genomes project subsequently undertaken by Genomics England Ltd (Department of Health & Social Care, United Kingdom). Overall, 13,187 BRRD participants, recruited either with a rare disease or as an unaffected relative, underwent whole

genome sequencing performed by Illumina on behalf of the BRRD. Of these, 7,388 individuals were recruited to one of fifteen rare disease domains, including: Bleeding, Thrombotic and Platelet Disorders (BPD), Cerebral Small Vessel Disease (CSVD), Ehler-Danlos and Ehler-Danlos-like Syndromes (EDS), Hypertrophic Cardiomyopathy (HCM), Intrahepatic Cholestasis of Pregnancy (ICP), Inherited Retinal Disorders (IRD), Leber Hereditary Optic Neuropathy (LHON), Multiple Primary Malignant Tumours (MPMT), Neurological and Developmental Disorders (NDD), Neuropathic Pain Disorders (NPD), Pulmonary Arterial Hypertension (PAH), Primary Immune Disorders (PID), Primary Membranoproliferative Glomerulonephritis (PMG), Stem cell and Myeloid Disorders (SMD) and Steroid Resistant Nephrotic Syndrome (SRNS). Individuals recruited as part of a Genomics England Limited (GEL) pilot study were included (Figure 2.1).

Details outlining the recruitment process and cohort summary for HCM cases have previously been reported.[121] Briefly, between May 2014 and September 2016, 246 patients with a clinical diagnosis of HCM, provided informed consent to participate within the BRRD (REC reference: 13/EE/0325). Eligible individuals demonstrated a clinical diagnosis of HCM, without evidence of pathogenic or likely pathogenic variants according to ACMG/AMP criteria, across 13 well-established HCM genes (8 core sarcomere genes: *MYBPC3, MYH7, TNNI3, TNNT2, MYL2, MYL3, ACTC1, TPM1*; 2 phenocopy genes: *PRKAG2, GLA*; and 3 genes with rarer associations: *FHL1, CSRP3, PLN*) following diagnostic clinical genetic testing. Additionally, recruitment was directed towards individuals in whom familial transmission was suspected in an attempt to enrich for possible "novel" genetic signals. Individuals were deemed eligible if they were aged between 18 and 70 years old, or over 70 years old with a family history of HCM. Eligible individuals were recruited through specialist ICC clinics in the United Kingdom, specifically: Oxford University Hospitals NHS Foundation Trust, Royal Brompton & Harefield NHS Foundation Trust, Guy's and St Thomas' NHS Foundation Trust, and The Newcastle upon Tyne Hospitals NHS Foundation Trust.

Individuals enrolled in the BRRD through a rare disease other than HCM were considered as reference controls. Available phenotypic details, provided by the BRRD, were surveyed and individuals in whom there was suspicion of an underlying ICC were excluded. As genome-wide sequencing had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant.

2.2.5 Type two diabetes mellitus

Access to unaffected control groups for case-control sequencing studies can prove challenging. In an optimal study design, controls would be matched against cases and screened for any co-morbidities that may confound analyses, sequenced and bioinformatically processed alongside cases to avoid systematic differences. Such studies have been performed in HCM, but require a dedicated recruitment effort that was not possible during this study.[51] As a compromise, individuals who have been recruited for an alternative purpose may prove suitable as a control group.

Here, 12,297 individuals, identified as being of European ancestry and not closely related (at least 3 degrees separation) via principal components analysis, recruited for case-control analysis of T2DM were made available for the purposes of HCM case-control analyses.

Details regarding the T2DM cohort have been published previously.[122] The full T2DM cohort consists of a multi-ancestry collection of 20,791 T2DM cases and 24,440 controls. 12,297 individuals of European ancestry were specifically identified from the following consortia: GoT2D, T2D-GENE and LuCAMP.

T2DM is a common, genetically complex disease of late onset that does not clinically resemble sarcomeric HCM and is not anticipated to yield rare variants in cardiomyopathy associated genes. However, as individuals recruited to the T2DM cohort were not specifically phenotyped for evidence of cardiomyopathy it is theoretically possible that this cohort may include individuals with HCM. It is assumed that the maximum prevalence of HCM in the T2DM cohort is equivalent to that observed in the general population (i.e. 1 in 500 individuals). This would

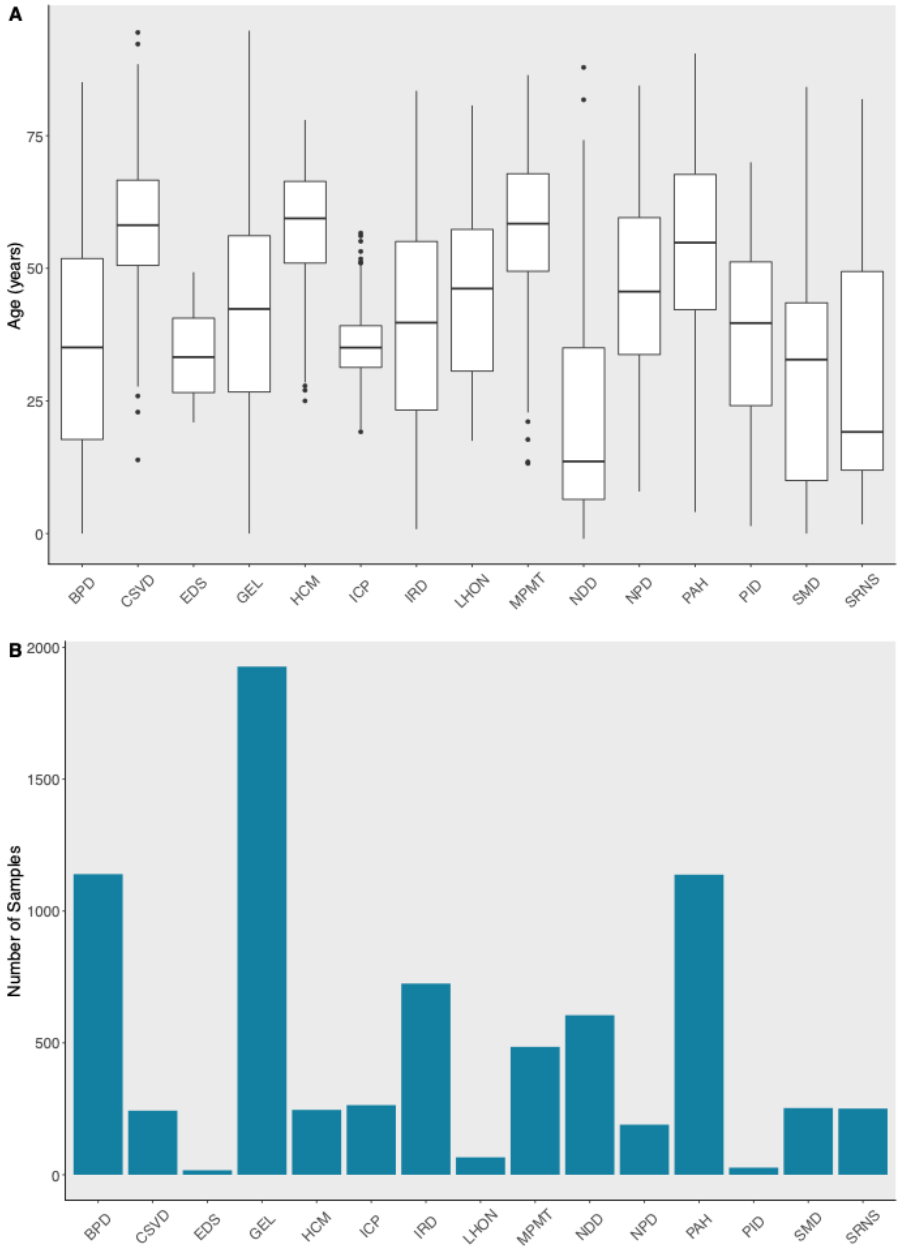


Figure 2.1: Summary of the BioResource for Rare Diseases. Panel A: Graphical representation of age distribution stratified by rare disease category. Panel B: Number of samples that contribute towards each rare disease category. Abbreviations: Bleeding, Thrombotic and Platelet Disorders (BPD), Cerebral Small Vessel Disease (CSVD), Ehler-Danlos and Ehler-Danlos-like Syndromes (EDS), Hypertrophic Cardiomyopathy (HCM), Intrahepatic Cholestasis of Pregnancy (ICP), Inherited Retinal Disorders (IRD), Leber Hereditary Optic Neuropathy (LHON), Multiple Primary Malignant Tumours (MPMT), Neurological and Developmental Disorders (NDD), Neuropathic Pain Disorders (NPD), Pulmonary Arterial Hypertension (PAH), Primary Immune Disorders (PID), Primary Membranoproliferative Glomerulonephritis (PMG), Stem cell and Myeloid Disorders (SMD), Steroid Resistant Nephrotic Syndrome (SRNS), and Genomics England Limited pilot (GEL).

correspond to a maximum of up to 25 individuals from the T2DM cohort (i.e. 1/500 multiplied by 12,297 individuals) with phenotypic evidence of HCM. Consequently, whilst the inclusion of cryptic HCM in the control groups may subtly influence test statistics, it is unlikely to be a major contributor towards possible type 2 errors (i.e. false negatives), and is comparable to what could be achieved through use of other available population controls. All individuals underwent exome sequencing, however only genetic regions captured by the 35 genes on the cardiomyopathy gene panel are available for interrogation. Access to these data were provided by the principal investigators of the T2DM consortia (Mark McCarthy, Jose Florez & Michael Boehnke). An advantage of using the T2DM cohort over large reference control datasets such as gnomAD was the ability to access individual level data. Therefore, whilst cognisant of the possible limitations, the T2DM cohort was selected as a reasonable control group for the purposes of rare variant analyses.

2.2.6 UK Biobank

The UK Biobank (UKBB) is a large population-based cohort study that consists of 502,543 individuals recruited from the United Kingdom, as described elsewhere.[123] Access was provided through UKBB application 11223. Individuals who expressed a wish to be withdrawn from the UK Biobank, as of 16/10/2018, were excluded.

Case-control status was determined using ICD10 codes (derived from HES data, self-reported questionnaire fields or death certificates) and rare variant data from exome sequencing. 386 cases were identified through ICD10 codes for HCM (I420 or I421). Individuals who possessed ICD10 codes for phenotypes that may confound analyses were excluded (n=15,901) from the control set (Table 2.1). For the ∼50k individuals who undertook the first tranche of exome sequencing, individuals were screened for variants of either uncertain significance, likely pathogenic or pathogenic status, as determined by the ACMG guidelines, across 35 cardiomyopathy-associated genes (*ACTC1, ACTN2, ANKRD1, BAG3, CRYAB, CSRP3, DES, DMD, DSC2, DSG2, DSP, FHL1, FHL2, FLNC, GLA, LAMP2, LMNA, MYBPC3, MYH7, MYL2, MYL3, PKP2, PLN, PRKAG2, RBM20, SCN5A, TMEM43, TNNC1,*

TNNI3, TNNT2, TPM1, TTN, TTR, TAZ, and *VCL*) and if present, excluded from the control set. For cases and controls, closely related individuals, within 3 degrees of relatedness, were excluded. The final control set included 270,260 individuals from which random samples (age and gender matched) were selected as an appropriate comparator group. As genome-wide genotyping had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant

2.2.7 Genomics England 100,000 Genomes Project

The Genomics England 100,000 Genomes Project (GeL) sequenced 101,162 genomes from 90,643 individuals recruited via 13 Genomic Medicine Centres across the United Kingdom. 73.8% of genomes sequenced (74,674 /101,162) were for individuals affected by a rare disease. Longitudinal and phenotypic information was provided via hospital episode statistics (HES) and clinician entered human phenotype ontology (HPO) terms. 952 individuals, of whom 811 were probands, were recruited based on a diagnosis of HCM. Access to the Genomics England data was provided via the Cardiovascular Genomics England Clinical Interpretation Partnership (GeCIP) Domain.

Analyses were performed on a subset of the total cohort, specifically 59,464 participants (release v5.1) mapped to GRCh38, identified by the GeL bioinformatic team based on inclusion criteria: $\geq$ 250 bp insert size, $\geq$ 75% mapped reads, $<$ 2% chimeric DNA fragments and <5% cross contamination. Kinship coefficients identified a group of 38,344 individuals who were not closely related who were subsequently used for downstream analyses. As genome-wide sequencing had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant.

2.2.8 Amsterdam Medical Centre

Details of the Amsterdam Medical Centre case series were provided by Dr Connie Bezzina (Amsterdam University Medical Centre) for the purposes of replication

for a genetic risk score. I had no involvement in either the recruitment or processing of patient data. 999 cases were identified using current diagnostic criteria (left ventricular wall thickness ≥15mm, or ≥13mm in presence of family history) from cardiovascular genetics referral centres in the Netherlands (Amsterdam University Medical Center, Erasmus Medical Center and the University Medical Center Groningen).[124] 2,117 controls were derived from a population cohort study from the Netherlands. As genome-wide genotyping had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant.

2.2.9 Royal Brompton Hospital

Details of the Royal Brompton case series were provided by Dr James Ware (Imperial College, London) for the purposes of replication for a genetic risk score. I had no involvement in either the recruitment or processing of patient data. 411 HCM cases were recruited from the Royal Brompton & Harefield Hospitals NHS Trust Cardiovascular Research Biobank. 1,211 controls, screened for evidence of HCM using cardiac imaging, were recruited from the UK Digital Heart Project.[125] As genome-wide genotyping had been performed, it was possible to undertake principal components analysis and genetically infer ancestry for each participant.

2.3 Reference data

2.3.1 Genome Aggregation Database

The Genome Aggregation Database (gnomAD v2.1) details the allele frequency of variants present in 141,456 individuals who previously undertook either exome (n=125,748) or genome (n=15,708) sequencing as part of a primary research study, investigating either population genetics or common, complex genetic disease. Individuals with severe, paediatric-onset disease, and their first degree-relatives, were excluded by the central gnomAD team. Participants explicitly consented to the sharing of their genotypes, in an aggregated format, in a publicly accessible database. Details of the gnomAD cohort are published elsewhere.[52]

Whilst gnomAD is not anticipated to be enriched for individuals diagnosed with HCM, their exclusion cannot be guaranteed as ICCs tend not to present in early childhood and no specific clinical phenotyping information is available for review. However, it is also plausible that the actual prevalence of HCM in gnomAD is lower than would be observed in an unscreened population, due the exclusion criteria applied to the constituent studies.

2.3.2 Trans-Omics for Precision Medicine

The Trans-Omics for Precision Medicine (TOPMed) Program, part of the Precision Medicine Initiative, contains participant data for over 149k participants encompassing more than 80 constituent studies, of varying study design, that have predominantly considered cardiac (39%), respiratory (33%), blood (8%) and sleep (1%) disorders.[54] Individuals with HCM were not actively recruited. Genome sequencing, with a median depth of 30X, was performed for a proportion of TOPMed participants using Illumina HiSeq X technology and is described in detail elsewhere.[54] Allele frequency data, derived from the genome sequencing of 62,784 TOPMed participants (freeze5), is publicly available via dbGAP and the BRAVO variant server (https://bravo.sph.umich.edu/freeze5/hg38/).

2.3.3 Combined gnomAD-TOPMED dataset

Allele counts derived from 198,517 non-overlapping individuals present in gnomAD or TOPMed were used as unscreened population controls. In the absence of individual level data for gnomAD or TOPMed, it is assumed that none of these control individuals carry more than one disease-associated allele. This assumption is based on evidence that HCM is an autosomal dominant condition with <1% of affected individuals demonstrating multiple pathogenic/likely pathogenic variants in case cohorts that have undertaken contemporary variant classification using the ACMG/AMP framework.[91] However, it is acknowledged that only having access to summary level data is a limitation.

2.4 DNA sequencing

2.4.1 Oxford Medical Genetics Laboratory

DNA was extracted and sequenced by employees of OMGL at the time of clinical referral, in accordance with clinical workflows between 2013 and 2018. Clinical-grade diagnostic sequencing was performed using Agilent Technologies' (Santa Clara, California, United States) HaloPlex Target Enrichment System (75bp, single index) using an Illumina MiSeq sequencing instrument.

2.4.2 Hypertrophic Cardiomyopathy Registry

DNA was extracted from whole blood samples for 2,684 individuals by employees of OMGL between 2013 to 2018. Target DNA intervals, identical to those defined by OMGL for clinical gene panel sequencing, were enriched using a custom-designed TruSeq Custom Amplicon (TSCA) protocol (250bp, dual indexed) (Illumina) before amplicon-based sequencing for 35 cardiomyopathy-associated genes was undertaken using the Illumina MiSeq platform.

2.4.3 BioResource for Rare Disease

13,187 individuals enrolled to the BRRD underwent genome sequencing, as reported elsewhere.[120] Sequencing and initial quality control (QC) were carried out by Illumina, having been contracted by the BRRD investigators. Briefly, genome sequencing was performed using the Illumina TruSeq DNA PCR-Free sample preparation kit and an Illumina HiSeq 2500 sequencer. Over the course of the project, 3 read lengths were used: 100bp (377 samples), 125bp (3,154 samples) and 150bp (9,656 samples). Illumina performed sample and data level QC and ensured at least 95% of the autosomal genome was sequenced to a depth of 15X. Reads were aligned to the Genome Reference Consortium human genome build 37 (GRCh37) using Isaac Genome Alignment Software (version 01.14), before SNVs and indels were called using a Bayesian framework, implemented by the Isaac variant caller (version 2.0.17).[126] The mean depth of the BRRD cohort was 45X (range from 34X to 72X). Following extensive quality control undertaken by the BRRD central

bioinformatics team, VCF files were made available for downstream analyses via the University of Cambridge's High Performance Computing Service.

2.4.4 Type 2 diabetes mellitus

Details relating to the exome sequencing performed for individuals enrolled within a large case-control study investigating the genetic aetiology of type 2 diabetes have previously been reported.[127] Sequencing (Agilent Sure Select and Illumina Rapid Capture) and QC were carried out by the T2DM consortia.

In brief, the exome sequencing data included in this cohort were derived from multiple sources, using multiple platforms, and systematically aggregated and joint called through the harmonisation of unmapped BAM files using GATK.[127–130] The mean depth coverage across the exome was reported as 40X. VCF files were prepared by the Senior Team Leader for the McCarthy Laboratory (Dr. Anubha Mahajan) and shared directly with me for the purposes of this thesis.

2.4.5 UKBB

The first tranche of exome sequencing, conducted by Regeneron/GlaxoSmithKline on behalf of the UK Biobank, was performed on 49,959 individuals, as previously reported.[131] A modified Integrated DNA Technologies xGen Exome Research Panel v1.0 was used to capture 19,396 genes (39 megabases of the human genome), with 20X coverage achieved for 94.6% of sites using the Illumina NovaSeq 6000 platform (75 bp paired-end reads). Bioinformatic processing, including variant calling, was performed using an automated analysis pipeline developed by DNAnexus.

2.4.6 Genomics England 100,000 Genomes Project

Genome sequencing was performed by Illumina. Following quality control, performed by GeL, a total of 677,003,512 variants aligned to GRCh38 were identified from 59,464 participants. Quality control included: $<5\%$ missingness per site, $\geq$ 10X coverage per site, $\geq$ 15 genotype quality (GQ) per site and failed allelic balance test < 0.25.

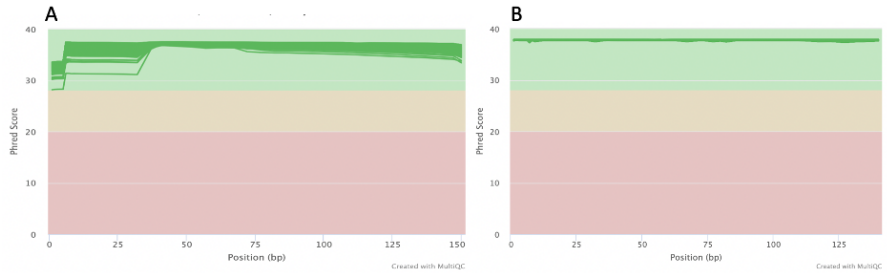


Figure 2.2: An example of MultiQC plots, generated from the HCMR cohort, and used to evaluate mean quality scores. Panel A: fastQC mean quality scores before trimming; Panel B: fastQC mean quality scores after trimming

2.5 Bioinformatic workflow

2.5.1 Pre-processing: HCMR and OMGL

FASTQ files specific to the OMGL and HCMR sequencing runs were retrieved and evaluated using MultiQC prior to quality control processing.[132] As the OMGL and HCMR FASTQ files were generated using differing sequencing approaches, initial pre-processing steps were performed specific to each cohort, with assistance from Silvia Salatino (High Throughput Bioinformatician, Wellcome Centre for Human Genetics, University of Oxford). For the OMGL cohort, reads were adapter-trimmed using Cutadapt, before restriction enzyme footprints were removed by cropping 5 base pairs from the 3' and 5' read ends to reduce the reference-bias that may have been introduced.[133, 134] For the HCMR cohort, target-specific primers and Illumina adapters were removed from the sequence data using Cutadapt. Subsequent steps in the bioinformatic pipeline were unified for the OMGL and HCMR cohorts.

Using Trimmomatic, reads were cropped at the transition point where the average Phred-scaled quality score was <30 (corresponding to a base call accuracy of 99.9%) within a 5 base pair sliding window.[135] Reads less than 50 base pairs in length were discarded. Read quality was subsequently evaluated using MultiQC (Figure 2.2).[132]

2.5.2 Variant calling: HCMR and OMGL cohorts

To limit heterogeneity between cohorts, filtered FASTQ files from both the OMGL and HCMR cohorts were processed together using a pipeline adapted from GATK Best Practices, using GATK version 4.0.11.0 and Picard version 2.9.2.[136, 137] GATK was selected for the purposes of variant calling due to its high sensitivity and specificity for germline variants, ability to generate a multi-sample VCF for an entire cohort, and standardised QC annotations to assist troubleshooting and comparison with independent datasets. A bed file was supplied to limit variant calling to genomic regions that had undertaken selective capture and amplification (Appendix A.2). Given the nature of amplicon-based sequencing, duplicates were deliberately left unmarked. Unmapped BAM files were generated from FASTQ files using GATK. Filtered reads were mapped onto GRCh37, specifically hs37d5, that includes decoy sequences to reduce false positive variant calling, using a Burrows-Wheeler Aligner (BWA-mem).[138] Unmapped BAM files were then merged with their corresponding BWA-aligned BAM files, co-ordinate sorted and indexed. GVCFs were generated using GATK's HaplotypeCaller, and consolidated via GATK's GenomicsDBImport function, before joint-calling was performed using GenotypeGVCFs. Variant quality score recalibration filtering was performed for single nucleotide variants (SNVs) and insertion-deletions (indels) separately. A truth sensitivity threshold of 99.9 and 99.7 was applied to SNVs and indels respectively using GATK's ApplyVQSR function, to reduce false positive variant detection. The resultant files were merged to generate a multi-sample VCF file.

2.5.3 Benchmarking using Genome in a Bottle resources

The performance of the OMGL and HCMR sequencing pipeline was evaluated using reference materials provided by The National Institute of Standards and Technology (NIST) hosted Genome in a Bottle (GIAB) Consortium (www.genomeinabottle.org). The GIAB Consortium have previously established a set of high-confidence genotype calls for DNA derived from an individual, NA12878, that has undergone extensive sequencing using over 13 different sequencing platforms and library preparation

techniques.[86] DNA derived from NA12878, extracted from a large homogenised growth of a B lymphoblastoid cell line, was sequenced and processed alongside OMGL and HCMR samples. Empirically generated variant calls for NA12878 were compared against the high-confidence genotype calls provided by the GIAB Consortium using the next generation sequencing (NGS) Benchmarking tool v1.4 (https://genomics.viapath.co.uk/benchmark), a tool supported by the Global Alliance for Genomics and Health (G4AH) Benchmarking Team.[88]

The sequencing pipeline designed to process the OMGL and HCMR samples performed well for SNVs and indels. Based on the sensitivity of each sequencing pipeline the OMGL/HCMR data outperformed a commercial sequencing pipeline (available from the Wellcome Centre for Human Genetics, University of Oxford), but was worse than the clinical grade sequencing pipeline measured across all available gene panels by OMGL, most notably for the detection of indels (Table 2.2). However, directly comparing each sequencing pipeline is challenging, given differences in the total number of variants that are captured by each sequencing pipeline. With the OMGL/HCMR pipeline capturing the fewest variants overall, small differences in marginal counts appear amplified relative to the WCHG exome or OMGL reference pipeline.

Additionally, the bioinformatic pipelines used to process the samples demonstrate differences. Critically, the OMGL/HCMR bioinformatic pipeline developed for this research has been developed and optimised for cohort-based analyses, and contrasts with the OMGL reference pipeline that was fine-tuned to process individual patient samples. As the OMGL reference pipeline was developed for clinical purposes, it preferentially detected false positives, rather than false negatives. The rationale being that false positives can be further evaluated (i.e. assessment of the underlying BAM files and/or confirmatory Sanger sequencing), whereas a false negative would never be detected and could impact patient care. This clinically focused approach differs from the cohort approach that was selected for the OMGL/HCMR bioinformatic pipeline (i.e. preference for false negatives over false positives). Consequently the cohort based approach may result in a slightly

lower pathogenicity yield than might be expected, but with the reassurance that the positive predictive value has been optimised. Overall, the OMGL/HCMR pipeline appears appropriate for the purposes of this cohort analysis.

Cohort	Variant type	Sensitivity	PPV	Total variants	TP	FN	FP
OMGL	SNPs	93.06%	100.00%	72	67	5	0
	Indels	66.67%	100.00%	2	2	1	0
HCMR	SNPs	93.06%	98.53%	72	67	5	1
	Indels	66.67%	100.00%	2	2	1	0
WCHG exome (Reference)	SNPs	90.16%	98.88%	56,395	47,289	5,162	537
	Indels	52.42%	93.04%	4,131	3,048	2,767	229
OMGL reference (>30X coverage)	SNPs	100%	98.9%	705	705	0	8
	Indels	96.3%	76.5%	27	26	1	8

Table 2.2: Genome in a bottle performance Comparison between OMGL, HCMR and reference series provided by the Wellcome Centre for Human Genetics and across all gene panels performed by OMGL. Abbreviations: PPV: positive predictive value; TP: true positive; FN: false negative; FP: false positive.

2.5.4 Post-processing quality control: HCMR and OMGL

Cohort specific VCF files were generated for HCMR and OMGL. Samples demonstrating poor coverage (defined as median coverage <100, a high proportion of missingness (>5%) or low transition-transversion ratio (tstv <2.13)) were removed. Genomic intervals, spanning a 211 kilobase region, were selected to correspond with the capture regions of the cardiomyopathy gene panel sequencing chemistries, as implemented by OMGL.

To provide a summary of the overall coverage achieved by the OMGL and HCMR pipelines, mean coverage estimates were generated for each cardiomyopathy associated gene using samtools (http://www.htslib.org/). Gene panel target capture intervals do not continuously span the entire gene and are instead enriched for coding sequence regions. Consequently, when comparing coverage between experiments, and reference coverage data from gnomAD, co-ordinates were synchronised, and

regions not included in the gene panel bed file discarded. This allows for continuous interpretation of coverage across captured regions of the gene. Whilst the sequencing methodologies are not directly compatible, between amplicon-based technologies and exome sequencing, the purpose of the comparison was to demonstrate that only a small proportion of captured regions demonstrate low coverage (i.e.<30X) (Figure 2.3).

Post-processing variant quality control was performed using empirically derived filter thresholds and visual interrogation of possible false positive variants of BAM files via the Integrative Genomics Viewer (IGV) Browser. Variant quality control filters were applied separately for SNVs and indels. SNVs were removed based on criteria recommended by GATK Best Practice Guidelines that assesses: variant quality by depth, corrected for allele depth (Quality by depth (QD) <5); mapping quality (MappingQuality (MQ) <40, or MappingQualityRankSumTest (MQRankSum) <-12.5); and bias, specific to both strand bias (FisherStrand (FS) >60) and position bias (ReadPosRankSum <-8). The distribution of true positive variants for SNVs and indels differs substantially. This largely reflects differences in mapping quality, as indels are of variable size they are penalised during alignment scoring processes. Consequently, filtering steps account for these systematic differences between variant classes to limit the exclusion of true positive variants. For indels, GATK Best Practice Guideline recommendations were followed and the following filtering criteria applied: QD <5, FS >200 or ReadPosRankSum <-20.

Across all variant classes, only variants achieving PASS filter status were retained, with variants demonstrating: an allele count <1, genotype depth <30X, genotype quality <30Q or overall quality score <30Q removed. Genotypes with sample-specific alternate allele read depths of <0.2 were removed. Variants with >5% missingness were identified and removed. An experienced OMGL employee (Michael Bowman) visually assessed each variant, via the Integrative Genomics Viewer (IGV), for common error modes to help identify and remove false positive variants, particularly within regions of low coverage, as would be performed for a clinical

diagnostic sequencing pipeline. A list of false positive variants identified through this process are listed in Appendix A.3.

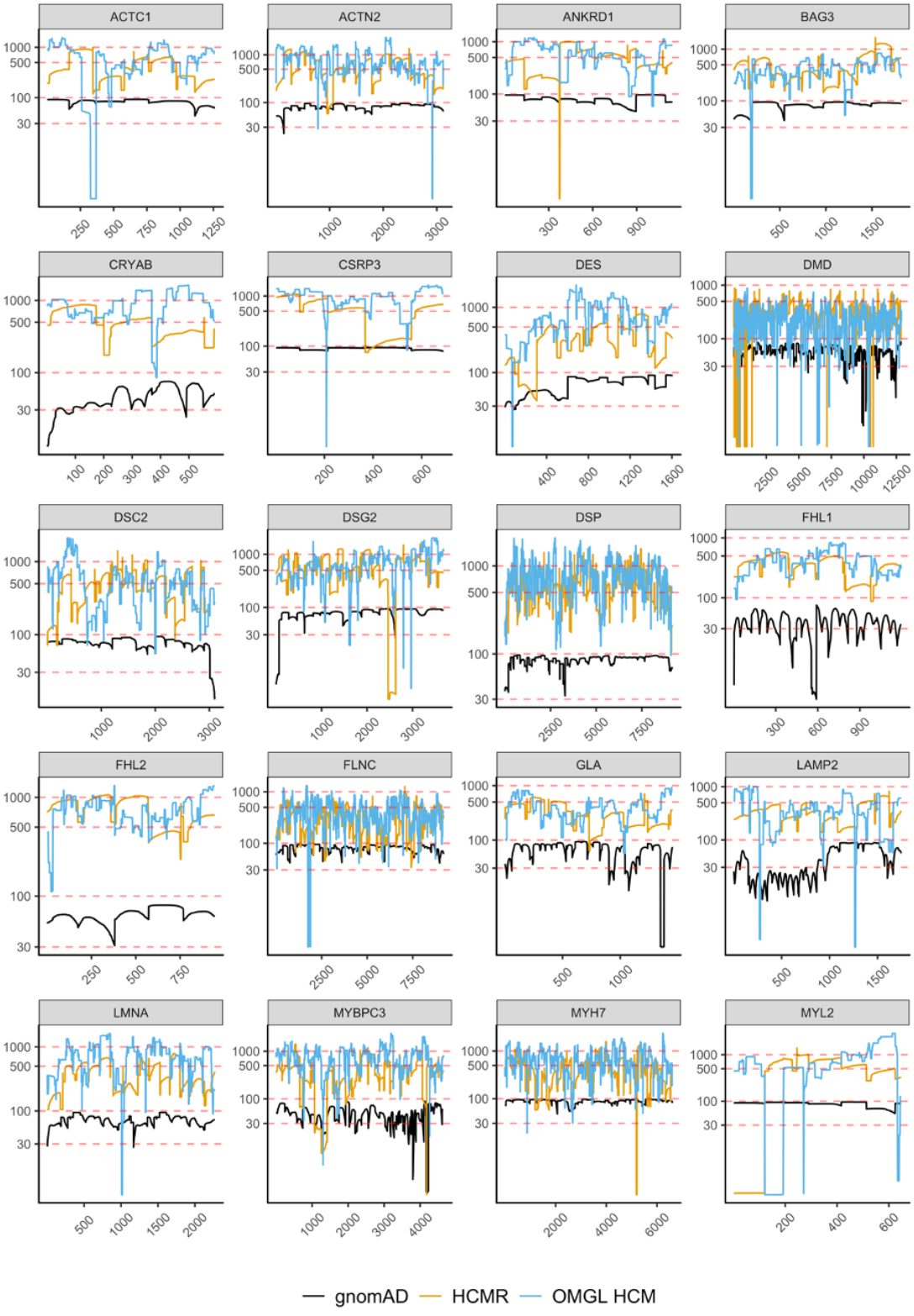
2.5.5 Sequencing quality control

Each cohort undertook a separate sequencing approach that introduced heterogeneity (Table 2.3).

Cohort	Individuals (n)	Depth (mean (sd))	Sites (mean (sd))	Proportion missing (mean (sd))	TsTv (mean (sd))	Sequencing platform
HCMR	2,636	57.1 (2.5)	5117 (0)	0.00714 (0.0177)	3.7 (0.74)	35 gene panel sequencing
OMGL HCM (2013-5)	1,017	49.9 (2.72)	4175 (43.2)	0.00893 (0.0121)	5.76 (1.94)	27 gene panel sequencing
OMGL HCM (2015-8)	1,740	61.1 (3.83)	4825 (0)	0.00564 (0.00366)	3.86 (0.803)	35 gene panel sequencing
BRRD cases	214	22.1 (2.02)	12688 (0)	0.0108 (0.00997)	3.12 (0.379)	Genome sequencing
BRRD controls	5,802	22.3 (2.33)	12688 (0)	0.0164 (0.0132)	3.15 (0.415)	Genome sequencing
T2DM	12,297	45.2 (12.3)	28207 (0)	0.0105 (0.00788)	4.11 (0.858)	Exome sequencing

Table 2.3: Quality control overview per cohort Overview of per-cohort sequencing quality and type of sequencing platform applied.

Mean depth per cohort values correspond to expected values derived for each sequencing instrument. The large variability in depth reported by the T2DM cohort reflects the composite nature of consortia efforts that have aggregated various exome sequencing methodologies. The mean coverage across all cardiomyopathy genes appears satisfactory, in both the OMGL and HCMR cohorts, and in gnomAD reference controls (see Figure 2.3). The BRRD cohort, which undertook genome sequencing, appears as an outlier with respect to coverage, but benefits from cases and controls being sequenced on the same platform. Quality control was also evaluated at an individual level across all available cohorts (Figure 2.4). To account for the addition of eight genes in 2015 to the OMGL gene panel, here OMGL results are reported for the pre (2013-2015) and post (2015-2018) this amendment. There


ACTC1
ACTN2
ANKRD1
BAG3
CRYAB
CSRP3
DES
DMD
DSC2
DSG2
DSP
FHL1
FHL2
FLNC
GLA
LAMP2
LMNA
MYBPC3
MYH7
MYL2
gnomAD
HCMR
OMGL HCM

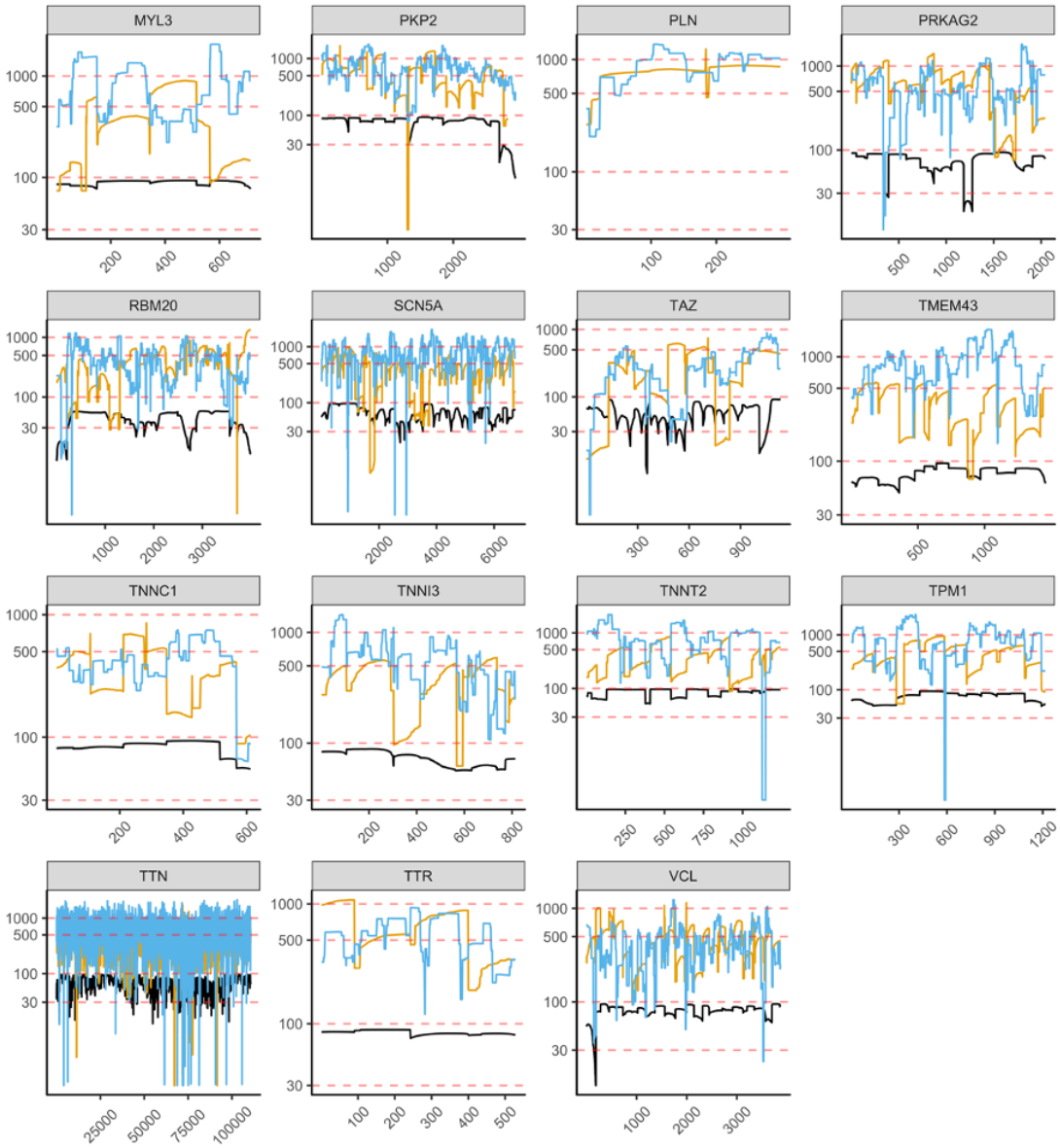


Figure 2.3: Coverage across 35 cardiomyopathy associated genes Mean coverage for 35 genes sequenced across HCMR and OMGL HCM cohorts and a reference cohort (gnomAD). The x-axis represents the length across the coding sequence for each gene and the y-axis represents the depth of coverage. Variants in regions <30X were removed. Coverage was only calculated for cohorts where BAM files were readily available for review (i.e. HCMR and OMGL cohorts). Coverage statistics were provided by gnomAD for comparison.

was considerable heterogeneity between sequencing methods. Depth between the cohorts cannot be easily compared due to differences in sequencing technology. The HCMR and OMGL cohort underwent amplicon based sequencing which is dependent on the creation of duplicated, non-independent reads, which contrasts with exome (T2DM) and genome (BRRD) sequencing that is dependent on the creation of independent reads. The transition-to-transversion (tstv) ratio appears inflated across the 35 cardiomyopathy genes and the relevance of this is uncertain. Whilst high tstv ratios could indicate an excess of artificially generated variants that would incur bias, it is also plausible that the inflated tstv ratios are a consequence of the gene panel sequencing capture methods being enriched for coding regions that contain high-GC content. For instance, tstv ratios are higher for exome sequencing experiments (3.0 - 3.3) than they are for genome sequencing (2.0 – 2.2) as a consequence of capturing a greater proportion of high-GC regions.[139, 140] The number of reported sites between the four cohorts differed despite identical genomic intervals being evaluated in each cohort. The HCMR and OMGL cohorts report ∼17% of the total number of sites documented for the T2DM cohort. The T2DM unrelated European data has been derived from a considerably larger multi-ancestry exome sequencing study. Consequently, all possible variants detected in the multi-ancestry VCF across the defined genomic interval, are detailed in the European-only subset, even if these variants do not specifically feature in the European subset of individuals. The HCMR cohort included individuals with relatively high rates of missingness. This is possibly attributable to the implementation of stringent QC performed at variant level, as variants that did not meet variant quality thresholds were replaced with missing values (see Table 2.4). As 31.9% (n=15/47) of individuals demonstrating >5% of missing sites harboured disease-causing variants (visually confirmed on BAM files), and other QC metric appearing satisfactory, it was decided to retain these individuals for subsequent analyses.

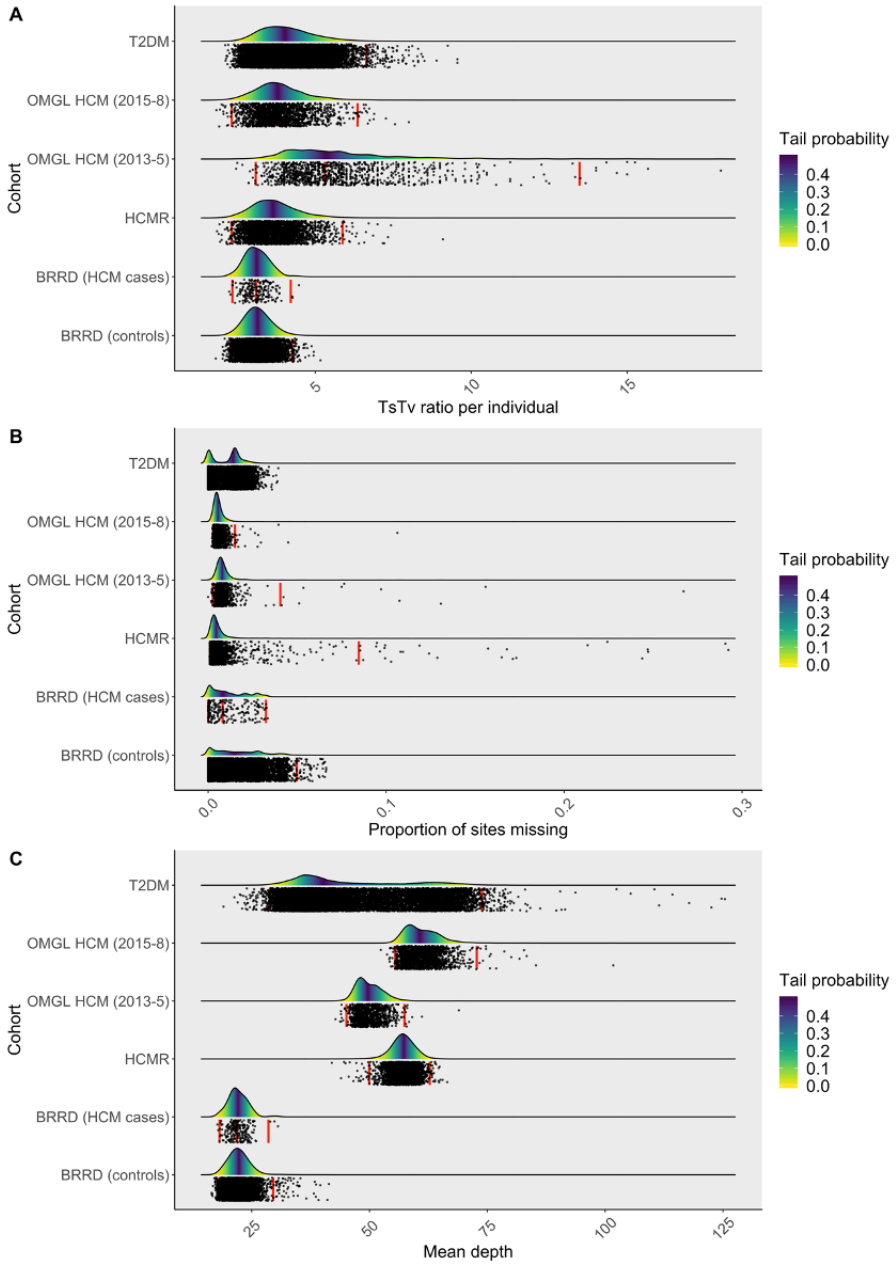


Figure 2.4: Individual level quality control summary of sequencing data provided across 35 cardiomyopathy associated genes. Each black dot represents a single individual. Red vertical lines represent the interquartile range and mean values. The distribution of individuals per metric is represented by as a probability density curve. The OMGL HCM (2013-15) cohort includes 28 genes, whilst all other cohorts evaluate 35 genes. Panel A) Ratio of transitions-to-transversions (TsTv) per individual. Panel B) Proportion of sites reporting missing information per individual. Panel C) Mean depth reported per individual across available genes.

2.6 Post-processing

2.6.1 Variant annotations

A unified annotation scheme was adopted for the OMGL, HCMR, BRRD and T2DM cohorts. Following extensive variant QC, variants were annotated using Ensembl's Variant Effect Predictor (VEP version 95), LOFTEE (https://github.com/konradjk/loftee), SNPEff and dbNSFP (version 4.0b2a).[52, 141–143] Additional annotations were included using bcftools (https://samtools.github.io/bcftools/bcftools.html). This included allele frequencies derived from publicly available resources, specifically gnomAD exomes and genomes (v2.1) and TOPMed freeze 5. Whilst gnomAD VCF files were mapped against GRCh37, TOPMed VCF files, were mapped against GRCh38, and following the download of these resources using the db-SNP portal (https://bravo.sph.umich.edu/freeze5/hg38/download), genomic co-ordinates corresponding to GRCh37 were annotated using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Variant classifications provided by ClinVar (version 20190211) were annotated alongside ACMG/AMP classifications derived from the OMGL mutation database.[91]

2.6.2 Gene and transcript selection

Multiple transcripts can originate from within a single protein coding gene and contribute towards a range of biological functions across multiple tissues. Traditionally, analyses have utilised transcripts derived from the longest coding region of a gene that are identified as "canonical" transcripts by reference databases, such as the NCBI Reference Sequence Database (RefSeq) and EMBL-EBI Ensembl database.[144, 145] However, the precise co-ordinates of the canonical transcript may differ between these reference resources, as they have been developed independently from one another and discrepancies between canonical transcript definitions are known to exist. Here, RefSeq transcripts, routinely interrogated by OMGL during clinical diagnostic workflows, were selected, as the sequencing chemistries and

capture methods are optimised for their evaluation. Ensembl transcripts have been annotated to enable direct comparison with additional resources using VEP.

It is, however, partially assumed that the "canonical" transcript present in the RefSeq database will have biological relevance in cardiac tissue. The advent of RNA sequencing has helped systematically evaluate and prioritise the relevance of each transcript across tissues of interest.[146] Of the genes included on the gene panel, RNA sequencing has proven most useful for the evaluation of variants in *TTN*, a gene that encodes the longest human protein, titin, critical to striated muscle function. Truncating variants in *TTN* account for between 15-20% of all DCM cases, with enrichment in the titin A-band.[147, 148] *TTN* undergoes extensive splicing and not all transcripts derived from TTN are present as isoforms in cardiac tissue. *TTN* is composed of 364 exons, however only those exons that contribute to more than 90% of transcripts in cardiac tissue (corresponding to percentage spliced in (PSI) score of >90%) are considered clinically informative for the purposes of identifying disease-causing variants.[148–150] Here, only high PSI exons are considered with respect to *TTN*.

2.6.3 Monogenic allele frequency thresholds

For the purposes of burden testing, an allele frequency filtering model that considers rare genetic variants based on the lower bound of a one-sided 95% confidence interval derived from a Poisson distribution, at an allele frequency of 1×10^{-4}, across specific populations (POPMAX) and across all available individuals in gnomAD was employed.[94]

Population-specific allele frequencies were determined using 125,748 exomes provided by gnomAD (v2.1) and total allele frequencies were determined using non-overlapping individuals from gnomAD v2.1 (exomes and genomes) and TOPMED freeze5 (genomes).[54, 90] The rationale for using an allele frequency threshold of 1×10^{-4} is derived from analysis performed by Whiffin et al (2018).[94] Assuming a

monogenic model of HCM, underpinned by an autosomal dominant inheritance pattern, the maximum credible allele frequency can be calculated using the prevalence, allelic heterogeneity and penetrance of HCM using the following formula:

$$Disease\ causing\ allele\ frequency_{max} = \frac{disease\ prevalence \times\ allelic\ heterogeneity}{penetrance\ \times\ 2}$$

Disease prevalence relates to the probability of disease within the general population at a fixed point in time; allelic heterogeneity denotes the probability an individual has a given genotype, conditioned on the presence of disease; and penetrance is the probability an individual develops a disease, conditioned on the presence of a given genotype. Given the diploid nature of the human genome, a constant of ½ is also applied.

Epidemiological studies have documented a clinical prevalence of 1:500 for HCM across multiple continents, using echocardiography and cardiac magnetic resonance imaging, with no extreme geographical differences. Allelic heterogeneity is a well-recognised feature of HCM, and prior case series have demonstrated that the most frequently observed pathogenic variant, *MYBPC3* c.1504C>T p.R502W, accounts for approximately 1.7% [95% CI: 1.4 - 2.0%] of disease. Variable penetrance is frequently observed in families affected by HCM. Establishing an accurate penetrance estimate for HCM has proved to be challenging; most estimates are derived from relatively small family-based studies that are susceptible to ascertainment bias, given that they are performed via an affected proband and are therefore enriched for genetic and/or environmental effects. Traditionally a penetrance of 0.5 has been denoted to reflect variable penetrance, but it is widely acknowledged to be the least reliable variable within this equation. Whilst the maximum credible allele frequency for HCM has been estimated as 4×10^{-5} (i.e. based on the 95% confidence interval upper bound for the allele frequency specific to the most frequently observed pathogenic variant in HCM (*MYBPC3* c.1504C>T p.R502W)), a slightly less stringent filtering allele frequency threshold of 1×10^{-4} remains routinely implemented, as this accommodates for a degree of uncertainty, allowing for penetrance estimates to be as low as

0.2. Whiffin et al (2018) also developed concepts outlined within the ACMG Guidelines regarding the impact of population-specific variants on allele frequency filtering thresholds, and introduced the concept of population-specific filtering.[94] The rationale for this being that if the prevalence of a disease is stable across geographical region, the allele frequency of a disease-causing variant should also be stable, presuming no ancestrally-specific modifiers exist.

2.6.4 Variant categorisation

Variants were categorised into groups based on properties annotated by VEP.[141] This approach is contingent on accurate mapping, variant calling and annotations. The variants were categorised as predicted loss-of-function, non-truncating and synonymous variants.

Although HCM is predominantly caused by missense variants, loss-of-function variants that result in haploinsufficiency in *MYBPC3* are a well-established cause disease. Loss-of-function variants were selected if the Loss-of-Function Transcript Effect Estimator (LOFTEE) plugin for VEP considered a variant to be of "high confidence".[52] LOFTEE employs a rules-based system to evaluate stop-gained, splice site disrupting and frameshift variants in an attempt to limit the inclusion of annotation errors. Non-truncating variants were included based on the following VEP annotated consequences: missense variant, inframe deletion, inframe insertion, stop lost, mature miRNA variant and start lost. As a group, synonymous variants are presumed to demonstrate a neutral effect on disease risk. Consequently, synonymous variants, identified using VEP annotated consequences, were used as a negative control for technical factors.

2.6.5 Variant classification

Guidelines published by the ACMG were used by employees of OMGL to classify variant pathogenicity into one of five discrete categorical variables (benign, likely benign, variant of uncertain significance, likely pathogenic, pathogenic) as part of a routine clinical workflow, between 2011 and 2019.[91] Rare variants (allele

frequency < 0.0001 in gnomAD) located in genes robustly associated with disease that had not previously been encountered by OMGL were identified based on in silico predictions regarding their impact on protein functionality via VEP, and classified using the ACMG/AMP variant classification framework.

2.6.6 Stratification based on sarcomere rare variant status

To dichotomise HCM cases into sarcomere-positive and sarcomere-negative groups VUS re-designation was performed by combining data from a range of sources, including: in-house and clinical variant databases, population databases (specifically gnomAD r2.1), gene-specific domain knowledge, and in silico prediction tools. This VUS re-classification was performed in collaboration with an experienced OMGL Clinical Scientist, Dr. Kate Thomson. A summary of this approach is outlined in Appendix A.4.

2.7 Genome-wide genotyping

2.7.1 HCMR

DNA was extracted from whole blood by an OMGL employee. The Oxford Genomics Centre (Wellcome Centre for Human Genetics, University of Oxford) received 2,784 genomic human DNA samples and processed samples according the workflows outlined in the Axiom 2.0 Assay Manual for the Precision Research Medicine Array (PMRA) (Affymetrix). 20µl of each sample was used, with DNA fragmented to between 25-125bp for genotyping purposes. No samples were removed prior to hybridisation. Samples were processed using 96-well plates and contained one Affymetrix control DNA sample per plate. Using the Axiom PMRA.r3 library from the Axiom Analysis Suite software, the overall failure rate was very low (0.07%); 1 sample failed to reach the Affymetrix recommended dish QC (DQC) threshold of 0.82 and 1 sample failed to reach the prespecified QC Call Rate threshold of 97%. The DQC threshold is a measure of genotyping performance that estimates the signal-to-noise ratio of non-polymorphic sites across the genome. The QC Call Rate provides an approximate estimate of sample quality based on a clustering call rate

algorithm, here provided by the Axiom GT1 algorithm for the PMRA. The overall call rate across the HCMR cohort exceeded 99.7%. Raw genotypes were transferred onto the University of Oxford's Biomedical Research Computing facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre for downstream analysis. Following the removal of duplicated samples (n=86), a total of 2,698 individuals were available for subsequent analysis.

2.7.2 BRRD

Using genome-sequence data, high-quality genome-wide variants were extracted from VCF files for 239 cases and 7,203 controls. High quality variants were defined as those that had; PASS filter status, a depth of at least 10 informative reads per site (DP > 10), a genotype quality score of at least 20 (GQ > 20), and a genotype missingness of no more than 10% (CR > 0.9). For the purposes of genome-wide association analysis, a priori power calculations indicated limited statistical power to detect rare variant disease associations (see chapter 6). Consequently, only high-quality variants with a minor allele frequency threshold of greater than 1% (MAF >0.01) were retained for analysis. Multiallelic sites were split, before variants were left-aligned and reference bases confirmed against genome build hs37d5 using bcftools. As genotypes were directly derived from genome sequencing data imputation was not necessary.

2.7.3 UKBB

438,427 UK Biobank participants underwent genome-wide genotyping using the Affymetrix UK Biobank Axiom® array, across 825,927 variants before extensive variant and sample level quality control was performed by the UKBB central team.[123]

2.7.4 Amsterdam Medical Center

Genotyping was performed using the Illumina Infinium BeadChip, Illumina Omni-Express and Global Screening Array. SNPs were mapped to GRCh37, and removed if: missingness rate >5%, HWE test p-value $<1\times10^{-6}$ for controls or p-value

$<1\times10^{-10}$ for cases, or MAF <0.05. Individuals were excluded when: missingness exceeded 3%, inbreeding coefficient ≥0.1, genotype-phenotype sex mismatch existed, proportional identity by descent >0.05, or non-European ancestry was indicated by principal components analysis. Phasing (Eagle2) and imputation (Haplotype reference consortium (HRCr1.1) panel) was performed on the Michigan Imputation Server v.1.0.2. SNPs with MAF>0.01 and a Minimac R^2>0.5 were retained.

2.7.5 Royal Brompton Hospital

Genotyping was performed using the Illumina Human OmniExpress beadchip. SNPs were mapped to GRCh37 and excluded if MAF<0.01, HWE P $<1\times10^{-7}$, or missingness rate >0.05. Sample QC excluded samples with a genotype-phenotype sex mismatch, heterozygosity rate >3 standard deviations from the mean, missingness rate >0.03 or evidence of non-European ancestry via principal components. Genotypes were phased using SHAPEIT (v2.r790) and imputed using IMPUTE2 (v2.3.2), against the UK10K and 1000 Genomes Project reference panel. SNPs with MAF >0.01 and INFO score >0.4 were retained.

2.7.6 Quality Control

EasyQC (v9.2) was implemented to perform quality control for each cohort, with the Haplotype Reference Consortium (HRC) reference panel used as reference material for mapping and allele frequencies.[151, 152] Variants were removed if they were monomorphic, demonstrated a minor allele count <6, were invalid, mismatched or duplicated, and when the observed allele frequency deviated by >0.2 from the HRC reported allele frequency.

2.8 Sample level quality control

Sample level QC was performed with consideration for ancestry, relatedness and gender ambiguity.

2.8.1 Principal components analysis to assign ancestry

Principal components analysis (PCA) is a dimensionality reduction technique that detects major sources of variation with a multivariate dataset. In population genetics, PCA was introduced by Price et al (2006) in an attempt to control for population stratification in case-control association studies, having previously been adopted to study geographical differences in population allele frequencies.[153–155] PCA is now established as a routinely implemented QC procedure.

Using a selection of genome-wide markers, not in linkage disequilibrium (LD) with one another or from an extended region of high LD (such as the human leukocyte antigen (HLA) region on chromosome 6), PCA generates a series of orthogonal principal components that are representative of features underlying variation in genome-wide structure, but may also detect other sources of variance including those relating to cryptic relatedness or technical artefact. The first principal component accounts for the largest proportion of variation, and each subsequent principal component explains a smaller proportion of variation.

Ancestral groups can be assigned using PCA by computing principal components for a reference dataset, such as the 1000 Genomes Project where recruitment was directed towards specific ancestral groups, and then onto this projecting experimental data to infer ancestral patterns.[156]

OMGL

No genome-wide SNP information was available and for anonymity reasons, self-identified ancestry information was not available. Using common variants (MAF > 0.1) in approximate linkage equilibrium with one another (PLINK function: –indep 50 5 1.5) located in the 211 kb region captured by gene panel sequencing, PCA was attempted but failed to satisfactorily discriminate between ancestral groups (Figure 2.5).

Consequently, I attempted to predict whether individuals were Non-Finnish European or not, based on the computed principal components. Non-Finnish European status was modelled using binary logistic regression, that incorporated

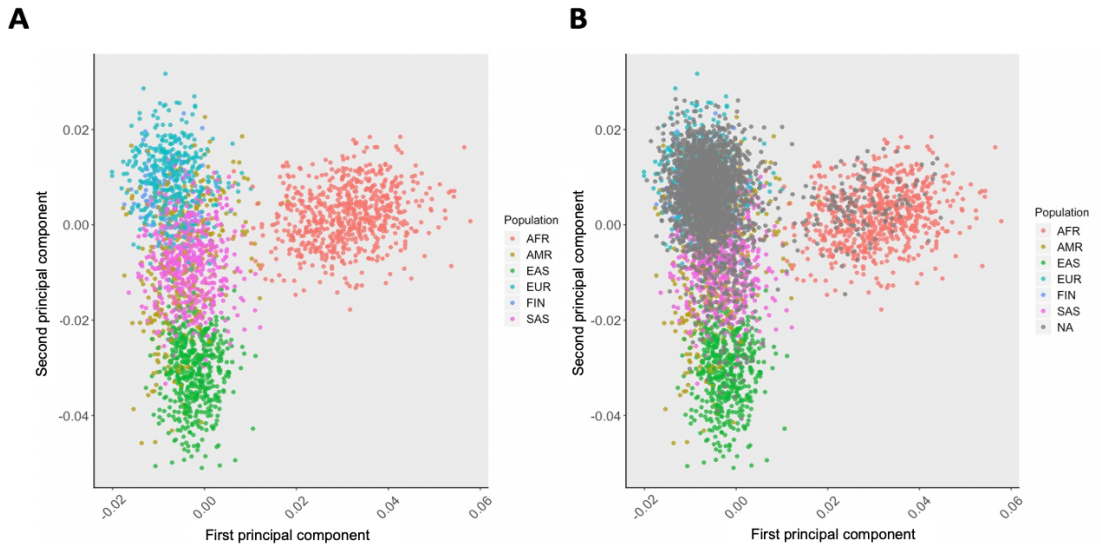


Figure 2.5: Principal component analysis performed using OMGL gene panel sequence data. Panel A: Common SNPs (MAF >0.1) present in 1000 genomes phase 3 data across 211kb region captured by OMGL gene panel used to partition ancestral groups. Panel C: Grey dots (coded NA) represent OMGL HCM samples, projected onto 1000 genomes phase 3 data. The majority of dots associate with European individuals. Abbreviations: AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian

the first 20 principal components with training data from the 1000 Genomes (20 principal components were selected to try to optimise the accuracy of the model, but I appreciate a similar result is obtainable with the first 5 principal components). Youden's J statistic was calculated, which gave a probability threshold of 0.142 and when applied to the OMGL cohort indicated that 1,980 individuals (78.3%), across 35 cardiomyopathy genes, were on average, of European ancestry.[157] However, it is acknowledged that the Non-Finnish European samples were effectively under sampled in the 1000 Genomes training data, as they represent ∼16% of the total cohort, and contrasts with the ∼80% expected in the OMGL data. In an attempt to account for this discrepancy a weighted model was designed and implemented using statistical software, SAS (Professor Martin Farrall), which accounts for the prior probability of being European in the OMGL dataset.[158] Implementation of a weighted model demonstrates improvements in the test sensitivity and specificity, as compared with the original logistic regression approach (Table 2.5). When

applied to the OMGL cohort, there is a slight increase in the number of individuals deemed to be European (n=2,024, 80%). Table 2.6 provides further information regarding the performance of the SAS weighted model when applied to both the 1000 Genomes Phase 3 data and the OMGL HCM cohort.

As no information was provided regarding self-identified ancestry, and it was not possible to perform genome-wide genotyping of ancestry informative markers, it was not possible to confirm these inferences.

HCMR

65,736 well-measured, genome-wide variants were used to conduct PCA, using flashPCA2, for 2,672 individuals from the HCMR cohort.[159] Variants used for PCA from the HCMR cohort were selected from 865,903 genotyped variants across 2,674 individuals. Exclusion criteria were applied, including: MAF <1% (n=388,695), Hardy-Weinberg equilibrium (HWE) exact p-value, with mid-p adjustment, $<1\times10^{-9}$ (n=14,397), genotyping call rate of <95% (n=5,020) or located in regions of high LD (chr5:44000000-51500000; chr6:25000000-33500000; chr8:8000000-12000000 and chr11:45000000-57000000) (n=10,888).[160] Two individuals were not included in the analysis as their genotyping call rate was <95%. 353,659 of the remaining 419,395 variants were removed following LD pruning with PLINK (using –indep-pairwise 1000 50 0.05, pairwise r^2 >0.05 with other variants located within a 1,000 variant sliding window).[161] FlashPCA2 was used to project ancestry-informative principal components present within the HCMR cohort onto the 1000 Genomes phase 3 cohort.[159, 161] Ancestry was predicted using a multinomial logistic regression model, from the nnet package (https://cran.r-project.org/web/packages/nnet/index.html), trained using ten ancestrally informative principal components derived from the 1000 Genomes Phase 3 data.[162] The predict function in R was then used to estimate ancestry, as determined by the International Genome Sample Resource (http://www.internationalgenome.org/category/population/) for HCMR individuals.[163] The kappa statistic, which incorporates the expected accuracy of the model compared with the observed accuracy, for the multinomial logistic regression

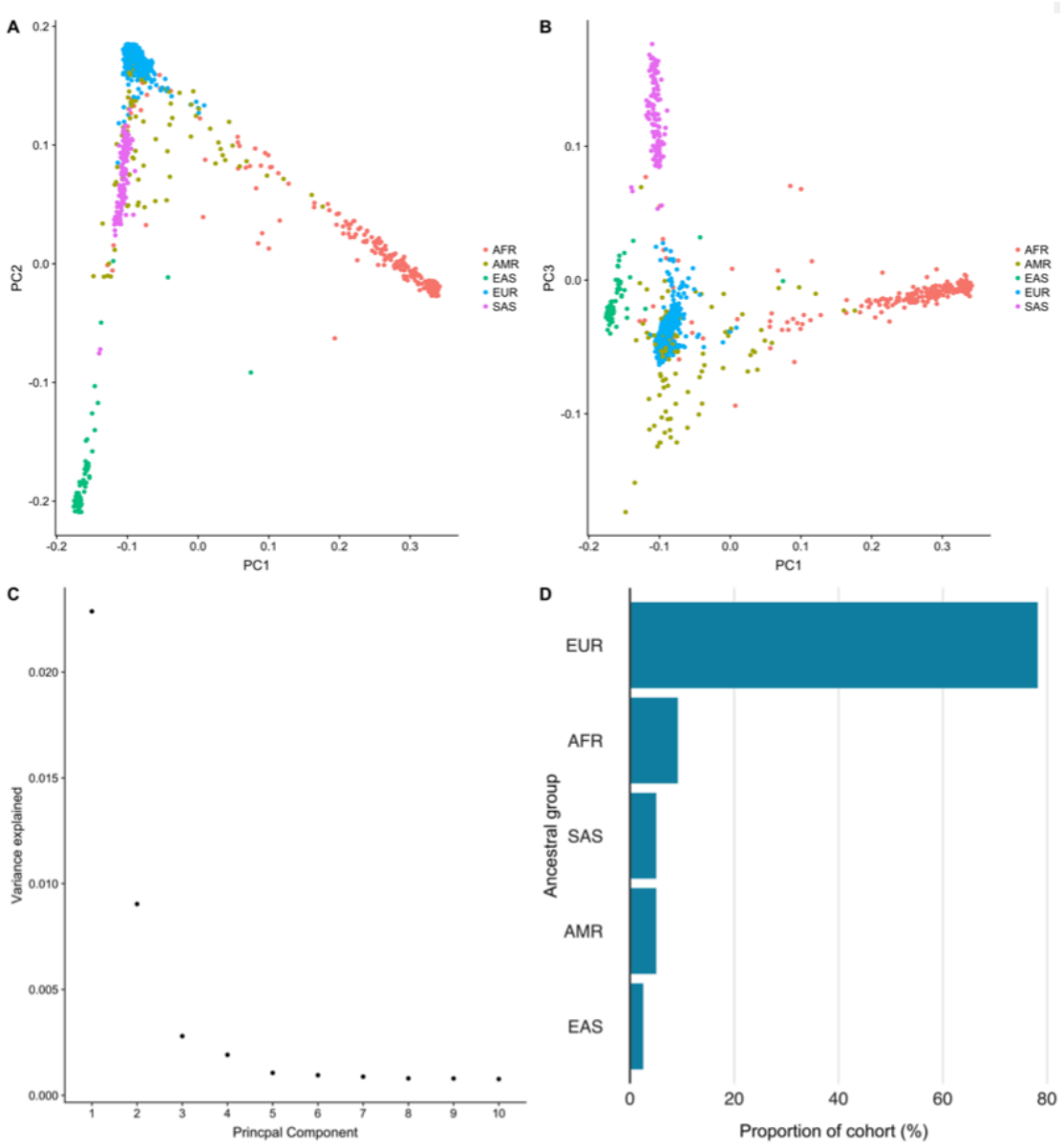


Figure 2.6: PCA of the HCMR cohort. Panel A: Principal component analysis evaluating ancestry (PC1 vs. PC2) within the HCMR cohort. Panel B: Principal component analysis evaluating ancestry (PC1 vs. PC3) within the HCMR cohort. Panel C: Scree plot representing the proportion of variation attributable to each principal component. Panel D: Relative proportion of ancestries contained within the HCMR cohort.

model was 99.9% [95% CI: 99.7 – 100.0%].[164] Summary statistics evaluating the performance of the multinomial logistic regression model are presented in Table 2.7 and Figure 2.6 presents results from the principal components analysis.

BRRD

Analysis was performed by the BRRD bioinformatics core team as previously reported.[120] Briefly, a set of 32,875 well-measured, unlinked (r^2 <0.2) and common (MAF > 0.3) SNPs were established using PLINK(v1.9), through which ancestry was assigned.[161]

Leveraging established ancestry groups assigned by the 1000 Genomes Project Consortium, principal components analysis (using the first 5 principal components) was performed to assign European, Finnish, African, East-Asian, South-Asian or Other ancestry, to individuals within the BRRD (Figure 2.7).

T2DM

PCA was performed by the T2DM consortia as previously described.[122] Briefly, common SNPs (MAF >1%), that demonstrated a 95% genotype call rate and were located further than 250Kb from either the HLA region or an established T2DM association signal underwent linkage disequilibrium pruning using PLINK (–indep pairwise 50 5 0.2). This generated a list of 171k variants from which the top 10 principal components of genetic ancestry were derived using EIGENSTRAT.[153] Only individuals of European ancestry were provided by the T2DM consortia.

2.8.2 Relatedness

The identification of closely related individuals in case-control genetic association studies was a critical QC procedure prior to the introduction of linear mixed models that can accommodate for this correlated structure in genome-wide association analyses. However, for burden testing that does not use linear mixed models, kinship coefficients were calculated wherever possible, and closely related individuals removed (3 degrees of relatedness), in an attempt to limit false positive associations.

OMGL

Without genome-wide genotyping information, relatedness could not be reliably inferred for this cohort and no relevant NHS record data were available for review,

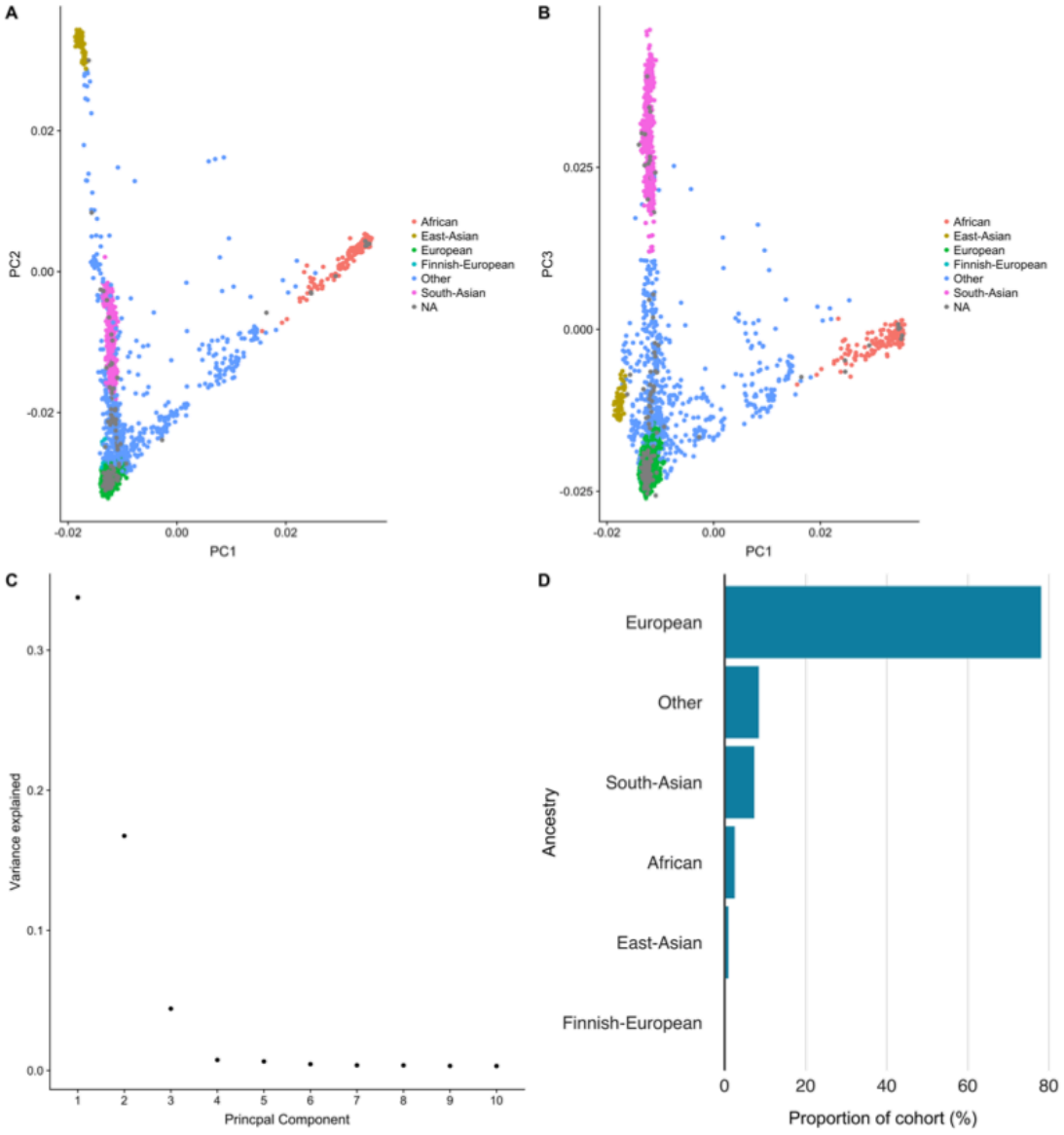


Figure 2.7: PCA of the BRRD cohort. Panel A: Principal component analysis evaluating ancestry (PC1 vs. PC2) in the BRRD cohort. Panel B: Principal component analysis evaluating ancestry (PC1 vs. PC3) within the BRRD cohort. Panel C: Scree plot representing the proportion of variation attributable to each principal component. Panel D: Relative proportion of ancestries contained within the BRRD cohort.

due to anonymity. Previous analysis performed using the OMGL cohort presumed probands were unrelated. This is reasonable given the clinical referral patterns, and necessary as the inclusion of related individuals can inflate both type 1 and type 2 errors.

HCMR

Using 730,064 genotyped variants from autosomes found on 2674 individuals, pairwise kinship coefficients were derived using KING (version 2.1.6).[165] Kinship coefficients were derived using identical by descent segment analysis, across 2672.3 Mb from 40 chromosomal segments. Relatedness was categorised using established kinship coefficient thresholds implemented by KING, specifically: >0.354 (duplicate/monozygotic twin), between 0.177 and 0.354 indicated 1st-degree relatedness, between 0.0884 and 0.177 indicated 2nd-degree relationships and 0.0442 to 0.0884 3rd degree relationships.

BRRD

Analysis was performed by the BRRD bioinformatic core team. Using PLINK (v1.9), a set of 32,875 well-measured, unlinked ($r2 < 0.2$) and common (MAF > 0.3) SNPs were generated. The principle components, derived from non-admixed individuals within the 1000 Genomes Phase 3 data, were used to generate a kinship matrix that accounted for population structure and provided a final set of unrelated individuals, based on kinship coefficients > 0.09. This was performed using KING, PC-AiR and PRIMUS.[165–167]

T2DM

It is reported that only distantly related individuals are included in the T2DM cohort (>3 degrees of relatedness).[122] Analysis was performed by the T2DM consortia and no kinship coefficients were shared.

2.8.3 Genetically derived gender

OMGL

Five genes (*DMD, GLA, LAMP2, FHL1* and *TAZ*) included on the OMGL cardiomyopathy gene panel are located on the X chromosome. 631 high-quality markers, not in linkage disequilibrium with one another, located on the X chromosome were present on 2,539 individuals (92.1%) from the OMGL cohort. Using an F statistic cut off of 0.8, PLINK suggested that 60.6% of the cohort were male (1533 males and 997 females). Unfortunately, it was not possible to validate these inferences with data provided on the clinical referral, due to concerns expressed by OMGL regarding confidentiality. Consequently, gender was taken as listed on the clinical referral and implies that 68.4% of the cohort were male.

HCMR

Self-reported gender was compared with genetically inferred gender, calculated using PLINK (Figure 2.8). Firstly, the pseudo autosomal region within the X chromosome was discarded before the –impute-sex function was employed. Male sex was characterised by an F statistic >0.8, and female gender was determined by an F statistic of <0.4 (Figure 8). Eleven individuals self-reported to be female were genetically estimated to be male. Similarly, six individuals self-reported as male were genetically estimated to be female. Individuals with discrepant genetic and self- reported gender assignments were discarded. For the BRRD and T2DM cohorts, gender analysis was performed by the respective core bioinformatic teams.

2.8.4 Statistical analysis

All statistical analysis was performed within the R computing environment (3.4.3) unless otherwise stated. Chi-squared tests were performed to compare differences between observed and expected rates. Confidence intervals were computed using binomial distributions, using Wilson's method.[168]

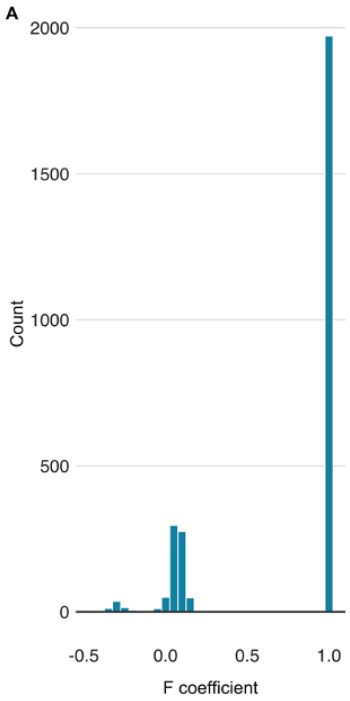


Figure 2.8: Summary of gender assignment in the HCMR cohort Assignment of genetically determined gender using the F-statistic. Male sex was characterised an F-statistic >0.8, and female gender was determined by an F-statistic of <0.4.

		Self-reported gender	
		Male	Female
Genetically estimated gender	Male	1923	11
	Female	6	758

Test statistics	
Accuracy	99.4% [95% CI: 99.0-99.6%]
Kappa	0.9845
Mcnemar's Test P-Value	0.332
Sensitivity	0.9969
Specificity	0.9857
PPV	0.9943
NPV	0.9921

Table 2.8: Summary of gender assignment in the HCMR cohort Summary of gender assignment between genetically determined and self-reported gender within the HCMR cohort

2.9 Discussion and limitations

This chapter documented the approach taken to generate appropriate materials to further evaluate the genetic aetiology of HCM.

Individual-level access to DNA sequence data enabled stringent quality control methodologies to be performed and comparison with external reference materials, including the United States National Institute of Standards and Technology's (NIST) Genome in a Bottle. This enabled genetic variation, present within the 35 cardiomyopathy associated gene capture region, to be formally evaluated. However, as this analysis is confined to data captured using amplicon-based sequencing across pre-specified genomic intervals, gene-discovery opportunities are relatively limited. Additionally, whilst amplicon-based sequencing is effective in densely sequencing prescribed regions, it is reliant on duplicate reads, which prevents direct comparison with other contemporary sequencing methods such as exome or genome sequencing that rely on independent reads, and thus coverage estimates cannot be directly compared between these two approaches.

It is important to appreciate that differences in sequencing technologies, between and within cohorts, may have introduced systematic bias. To safeguard against such effects, careful quality control has been undertaken. Documented in this Chapter are the results from the sample and variant level quality control procedures, which provide reassurance in the underlying data. It should be recognised that coverage files were not available for direct evaluation between all cohorts and this is a limitation to the quality control methods administered. Reassuringly, coverage for the OMGL and HCMR HCM cohorts were available, alongside summary level coverage statistics for gnomAD (Figure 2.3). Furthermore, there appears to be no systematic error or spurious result arising from using cohorts that did not have BAM files available for coverrage evaluation (i.e. the T2DM cohort or BRRD cohort).However, further quality control was also performed at a cohort level to evaluate for heterogeneity between cohorts, as outlined in subsequent Chapters. Acknowledgement of this potential limitation influenced cohort selection for specific experiments. For instance, in Chapter 4, the BRRD cohort was selected for the

evaluation of possible oligogenicity given that cases and controls were processed and sequenced in unison, reducing the risk of systematic differences between cases and controls. There are occasions throughout the thesis when case cohorts (i.e. the OMGL and HCMR cohorts) are combined to increase discovery power. In such scenarios, systematic differences between case cohorts were firstly assessed, and reported, to provide confidence in subsequent findings.

The HCMR cohort demonstrates several strengths over the OMGL cohort. The availability of genome-wide genotyping data has enabled detailed principal components analysis and relatedness to be performed. Furthermore, available clinical and demographic information has proven beneficial for the purposes of sample level quality control and will facilitate future experiments that require clinical correlation.

ICD10 code	Diagnosis
I460	Cardiac arrest with successful resuscitation
I461	Sudden cardiac death, so described
I469	Cardiac arrest, unspecified
I420	Dilated cardiomyopathy
I421	Obstructive hypertrophic cardiomyopathy
I422	Other hypertrophic cardiomyopathy
I423	Endomyocardial (eosinophilic) disease
I424	Endocardial fibroelastosis
I425	Other restrictive cardiomyopathy
I426	Alcoholic cardiomyopathy
I427	Cardiomyopathy due to drug and external agent
I428	Other cardiomyopathies
I429	Cardiomyopathy, unspecified
I430	Cardiomyopathy in infectious and parasitic diseases classified elsewhere
I431	Cardiomyopathy in metabolic diseases
I432	Cardiomyopathy in nutritional diseases
I438	Cardiomyopathy in other diseases classified elsewhere
I501	Left ventricular failure
I509	Heart failure, unspecified
I517	Cardiomegaly
I110	Hypertensive heart disease with (congestive) heart failure
I119	Hypertensive heart disease without (congestive) heart failure
I500	Congestive heart failure
I518	Other ill-defined heart diseases
I519	Heart disease, unspecified
I514	Myocarditis, unspecified
I515	Myocardial degeneration
I516	Cardiovascular disease, unspecified
O101	Pre-existing hypertensive heart disease complicating pregnancy, childbirth and the puerperium
O103	Pre-existing hypertensive heart and renal disease complicating pregnancy, childbirth and the puerperium

Table 2.1: ICD10 codes excluded from UKBB controls Individuals possessing an ICD code detailed in this table were excluded from the control set to limit possible confounding

		OMGL	HCMR
Sample QC	Total samples pre-QC	2,758	2,684
	Samples removed during QC	1	48
	Median <100X coverage	1	48
	Transition to transversion ratio <2.13	0	24
	Total samples post-QC:	**2,757**	**2,594**
Variant QC	Total variants pre-QC	11,624	11,624
	SNVs	10,855	10,855
	Indels	663	663
	Other	106	106
	Variants removed during QC:	6,783	6,376
	SNVs removed: QD < 5 or FS > 60 or MQ < 40 or MQRankSum < -12.5 or ReadPosRankSum < -8	1,759	1,759
	Indels removed: QD < 5 or FS > 200 or ReadPosRankSum < -20	122	122
	No PASS filter status	27	19
	Monomorphic sites	29	38
	Overall quality score < 30	9	10
	Genotype depth <30	2,562	2,414
	Genotype quality <30	2,510	2,128
	Variants with >5% missingness	90	293
	Total variants post-QC:	**4,841**	**5,248**
	SNVs	4,638	4,997
	Indels	171	205
	Other	32	46

Table 2.4: Quality control summary for HCMR and OMGL cohorts

	Standard approach	SAS weighted approach
Threshold	0.1423	Weighted 0.8000
Specificity	0.8552	0.8962
Sensitivity	0.9257	0.9455
Accuracy	0.8666	0.9042
True negative	1796	1882
True positive	374	382
False negative	30	22
False positive	304	218
Negative predictive value	0.9836	0.9884
Positive predictive value	0.5516	0.6367
Precision	0.5516	0.6367

Table 2.5: Performance of principal components analysis in OMGL Summary statistics of the logistic regression models used to estimate the genetic ancestry of the OMGL cohort. Trained using 1000 Genomes phase 3 data, and used to predict ancestry within the HCMR cohort. Standard approach involves a binary logistic regression model with 20 principal components as explanatory variables. The SAS weighted approach considers the prior probability of being European (p=0.8) in the OMGL dataset.

			Predicted ancestry		
	Population	Counts and percentages	Non-European	European	Total
Training data: 1000 Genomes Phase 3 data	AFR	n	660	1	661
		%	99.85	0.15	100
	AMR	n	246	101	347
		%	70.89	29.11	100
	EAS	n	504	0	504
		%	100	0	100
	EUR	n	22	382	404
		%	5.45	94.55	100
	FIN	n	21	78	99
		%	21.21	78.79	100
	SAS	n	451	38	489
		%	92.23	7.77	100
OMGL HCM		n	506	2,024	2,530
		%	20	80	100
	Total	n	2,410	2,624	5,034
		%	47.87	52.13	100

Table 2.6: Approximated OMGL ancestry Individuals partitioned based on the weighted SAS model into European or non-European. Presented training data from 1000 genomes phase 3 data and implementation in the OMGL HCM dataset Abbreviations: AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian.

	AFR	AMR	EAS	EUR	SAS
Sensitivity	0.9985	0.9971	1	1	1
Specificity	0.9995	0.9995	1	1	1
PPV	0.9985	0.9971	1	1	1
NPV	0.9995	0.9995	1	1	1
Prevalence	0.264	0.1386	0.2013	0.2009	0.1953
Detection Rate	0.2636	0.1382	0.2013	0.2009	0.1953
Detection Prevalence	0.264	0.1386	0.2013	0.2009	0.1953
Balanced Accuracy	0.999	0.9983	1	1	1
Total within HCMR (n, %)	245 (9.2)	135 (5.1)	68 (2.5)	2091 (78.2)	135 (5.1)

Table 2.7: Approximated HCMR ancestry Summary statistics of the multinomial logistic regression model for genetic ancestry, trained using 1000 genomes phase 3 data, and used to predict ancestry within the HCMR cohort Abbreviations: AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian.

3

Evaluating monogenic disease models

Contents

3.1 Background

The identification of genes causal of monogenic HCM benefited from several waves of technical innovation. Beginning with linkage-analysis over twenty years ago, the first cadre of HCM genes were identified, defining HCM as a disease of the sarcomere.[35–

41] However, alongside these now-venerated findings were other genes that emerged during an era, where, in retrospect, there were naïve assumptions regarding the abundance of rare genomic variation. Candidate gene studies emerged where novel variants, detected in HCM cases, were surveyed for in a few hundred controls; absence of a variant in the control cohort was deemed sufficient evidence to support causality. Examples of this include ankyrin repeat domain 1 (*ANKRD1*), myozenin 2 (*MYOZ2*), and troponin C1 (*TNNC1*).[169–171] Consequently, numerous genes established themselves on clinical genetic testing panels without robust evidence supporting causality. Most recently, the aggregation and harmonisation of large-scale sequencing experiments has characterised the null distribution and frequency of rare variation across the genome, facilitating a re-evaluation of genes, presumed to be causal of disease, through gene-level burden testing.[8, 53] As a result, a rationalised list of causal HCM genes emerged, predominated by HCM genes derived from genome-wide linkage analyses that involved multi-generational pedigrees. Not only has this improved our confidence in the genes deemed to be causal of HCM, but it has highlighted that disease-causing variants (pathogenic, likely pathogenic or variants of uncertain significance) are not detectable in the majority of individuals undergoing HCM genetic testing. Consequently, whilst a substantial proportion of heritable, monogenic HCM has been established, the remaining genetic architecture of HCM remains relatively unknown. As outlined in Chapter 1, several hypotheses emerge regarding the genetic architecture of HCM, including the possibility that additional genes contribute towards a monogenic model of HCM, but are yet to be formally elucidated.

The key objectives for this chapter are to:

1. Evaluate the monogenic architecture of HCM using available case series.

2. Perform rare variant burden analysis across a list of cardiomyopathy associated genes through a case-control analysis.

3.2 Monogenic architecture

Five case cohorts and four control cohorts were available for rare variant analyses, representing a total of 6,493 cases and 56,824 controls. The OMGL cohort represents a collection of probands, presumed to be unrelated and predominantly of European ancestry, who were referred for clinical diagnostic genetic testing to Oxford University Hospitals NHS Foundation Trust. The HCMR cohort included patients diagnosed with incident HCM across 44 clinical sites between Europe and North America. The BRRD cohort represents a rare disease pilot study that began recruitment prior to GeL. A specific exclusion criterion for individuals with HCM enrolled in the BRRD was the presence of a disease-causing variant across a panel of 14 core sarcomere genes. GeL is enriched for sarcomere-negative HCM, with recruitment directed towards individuals lacking a disease-causing variant on a routine gene-panel.

Three of the four control cohorts were sequenced alongside case data (BRRD, RBH, GeL). The T2DM cohort represents a heterogeneous collection of individuals, unscreened for HCM, who were originally recruited to a large T2DM consortia (Chapter 2).

3.2.1 Demographics

Demographic details for cases and controls are available in Tables 3.1 and 3.2.

The case cohorts demonstrated a preponderance towards male gender, aligning with prior reports.[55, 56, 172, 173] Given that HCM is an autosomal dominant condition, it is unexpected that HCM is more frequent in men. It is possible that this observation is attributable to diagnostic differences between men and women, as previous analysis indicates women with HCM present later in life with more advanced disease.[173, 174] However, it could also be hypothesised that these observed differences are reflective of underlying biological differences between men and women (i.e. differences in circulating sex hormone concentrations), and may offer insight into possible protective factors.

Where possible, genome-wide genotype data was used to infer relatedness and principal components analysis was performed to establish each individual's average

genetic ancestry. As the OMGL cohort lacked genotype array data relatedness and ancestry were not evaluated. Individuals recruited to the HCMR cohort were predominantly of European ancestry (78.3%); this reflects a more general trend towards an over-representation of individuals of European ancestry in biomedical research, including GWAS, that is not representative of global populations.[175]

	OMGL	HCMR	BRRD	RBH	GEL
Sample size	2,757	2,636	213	411	476
Age (years (SD))	54.5 (16.3)	49.5 (11.3)	58.6 (10.5)	65.7 (15.1)	55.7 (15.4)
Gender (male %)	68.4	71.4	80.8	71.0	69.5
Relatedness	Presumed not to be closely related. Unable to validate.	Closely related individuals removed (MZ: 5, PO: 28, FS: 44, 2nd: 16, 3rd: 49)	Closely related individuals removed	Closely related individuals presumed to have been removed	Closely related individuals removed
Ancestry (n, (%)):					
AFR	NA	239 (9.0%)	0 (0%)	NA	17 (3.6%)
AMR	NA	135 (5.1%)	0 (0%)	NA	1 (0.2%)
EAS	NA	68 (2.6%)	0 (0%)	NA	2 (0.4%)
EUR	2,206 (80.0%)*	2,074 (78.3%)	213 (100%)	411 (100%)	357 (75.0%)
SAS	NA	134 (5.1%)	0 (0%)	NA	64 (13.4%)
Not assigned	551 (20%)*	0 (0%)	0 (0%)	NA	35 (7.4%)
Variant carrier status:					
P	471 (17.1%)	572 (21.6%)	1 (0.47%)	48 (11.7%)	17 (3.57%)
LP	191 (6.9%)	216 (8.2%)	4 (1.88%)	44 (10.7%)	15 (3.15%)
VUS	392 (14.2%)	379 (14.3%)	18 (8.45%)	3 (0.7%)	26 (5.46%)
Negative	1,703 (61.8%)	1,483 (56.0%)	190 (89.2%)	316 (76.9%)	402 (84.4%)
Rare unclassified	0 (0%)	0 (0%)	0 (0%)	0 (0%)	16 (3.51%)
Countries samples derived from:	United Kingdom (2,166); New Zealand (221); Ireland (129); Sweden (125); Australia (39); Canada (17); Portugal (3); El Salvador (1); Cyprus (1); Brazil (1)	United States of America (1,121); United Kingdom (921); Canada (242); Germany (211); Netherland (153); Italy (115)	United Kingdom (213)	United Kingdom (411)	United Kingdom (476)

Table 3.1: Demographic summary of case cohorts Legend: OMGL: Oxford Medical Genetics Laboratory; HCMR: HCM Registry; BRRD: BioResource for Rare Disease; RBH: Royal Brompton Hospital; GEL: Genomics England 100k Genomes. MZ: Monozygotic twins/duplicates; PO: Parent offspring; FS: Full sib; 2nd: 2 degrees related; 3rd: 3 degrees related; Variant carrier status using ACMG classification for core sarcomeric genes (*MYBPC3, MYH7, TNNI3, TNNT2, MYL2, MYL3, ACTC1, TPM1*); Variant carrier status: P: pathogenic; LP: likely pathogenic; VUS: variant of uncertain significance. NA: Not available. BRRD cohort has been filtered to include only individuals of European ancestry, as ascribed by principal components projected by the 1000 Genomes Phase 3 data, with removal of closely related individuals. *80% of OMGL individuals are anticipated to be of European ancestry based on analysis performed in Chapter 2. The BRRD project encompasses multi-ancestry and related individuals, all of whom have undertaken genome sequencing and are used in other areas of this thesis. GEL ancestry determined by a probability exceeding 0.9 from a random forest model, based on the 1000 Genomes Phase 3 data.

	BRRD	T2DM	RBH	GEL
Sample size	5,801	12,297	1211	37,515
Age (years (SD))	46.0 (20.3)	58.5 (12.3)	47.0 (13.3)	48.3 (19.7)
Gender (male %)	40.2	50.9	46.1	45.8
Relatedness	Closely related individuals removed	Closely related individuals removed	Closely related individuals presumed to have been removed	Closely related individuals removed
Ancestry (n, (%)):				
AFR	NA	-	NA	1120 (2.99%)
AMR	NA	-	NA	92 (0.2%)
EAS	NA	-	NA	272 (0.7%)
EUR	5,801 (100%)	12,297 (100%)	NA	29621 (79.0%)
SAS	NA	-	NA	3080 (8.2%)
Not assigned	NA	-	NA	3330 (8.9%)
Variant carrier status:				
P	NA	NA	0 (0%)	84 (0.22%)
LP	NA	NA	0 (0%)	75 (0.20%)
VUS	NA	NA	0 (0%)	632 (1.68%)
Negative	NA	NA	1211 (100%)	35,477(94.6%)
Rare unclassified	NA	NA	0 (0%)	1247 (3.32%)
Countries samples derived from:	United Kingdom (5,801)	NA	United Kingdom (1,211)	United Kingdom (37,515)

Table 3.2: Demographic summary of control cohorts Legend: BRRD: BioResource for Rare Disease; T2DM: Type 2 diabetes mellitus consortia; RBH: Royal Brompton Hospital; GEL: Genomics England 100k Genomes. MZ: Monozygotic twins/duplicates; PO: Parent offspring; FS: Full sub; 2nd: 2 degrees related; 3rd: 3 degrees related; Variant carrier status using ACMG classification for core sarcomeric genes (*MYBPC3, MYH7, TNNI3, TNNT2, MYL2, MYL3, ACTC1, TPM1*); Variant carrier status: P: pathogenic; LP: likely pathogenic; VUS: variant of uncertain significance. NA: Not available. BRRD cohort has been filtered to include only individuals of European ancestry, as ascribed by principal components projected by the 1000 Genomes Phase 3 data, with removal of closely related individuals. *The BRRD project encompasses multi-ancestry and related individuals, all of whom have undertaken genome sequencing and are used in other areas of this thesis. GEL ancestry determined by a probability exceeding 0.9 from a random forest model, based on the 1000 Genomes Phase 3 data.

3.2.2 Pathogenicity yield

Individual level sequence data was available for the OMGL, HCMR, BRRD and GeL case cohorts and T2DM, BRRD and GeL control cohorts. Summary level statistics were made available for the RBH case and control cohorts. For the OMGL and HCMR cohorts a minimum of 27, and maximum of 35, cardiomyopathy associated genes underwent clinical-grade gene panel sequencing, performed in a United Kingdom Accreditation Service (UKAS)-accredited clinical diagnostic laboratory. Genome sequencing was performed for the BRRD and GeL cohorts

and individual level access was possible via a dedicated server. Whilst the T2DM cohort underwent exome sequencing, data was shared for 35 cardiomyopathy genes, corresponding to those regions sequenced using clinical-grade gene panel sequencing.

Inherent differences between the cohorts of cases were reflected by the overall pathogenicity yields. The HCMR cohort received a pathogenic or likely pathogenic variant classification more often than the OMGL cohort (788/2,636 (29.9%) vs. 608/2,757 (22.1%); p-value=6.14×10^{-11}). In subsequent analyses the OMGL and HCMR cohorts were combined. Overall, the OMGL-HCMR cohort demonstrated that 25.9% [95% CI: 24.7-27.1%] (n=1396/5,393) of individuals possessed either a pathogenic or likely pathogenic variant in at least one of eight core sarcomere genes (*MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3, TPM1* or *ACTC1*), and when expanded to incorporate both an expanded gene list (across a maximum of 35 cardiomyopathy associated genes) and the inclusion of variants of uncertain significance, the overall yield rose to 39.3% [95% CI: 38.0-40.6%] (n=2117/5,393). When compared with previously reported estimates (Alfares et al (2015): 32%; Walsh et al (2016): 32%), a pathogenic/likely pathogenic yield of 25.9% is lower than expected, but this may reflect advances in variant classification methodology and could be influenced by a circumscribed search space of eight core sarcomere genes.[8, 55, 93]

Despite the enrolment of sarcomere-negative HCM, re-analysis of the BRRD cohort, using genome-sequence data provided by the BRRD study and the OMGL variant classification catalogue, demonstrated that five individuals (2.35% [95% CI: 1.00-5.38%]) harboured either pathogenic (n=1, and *MYBPC3* p.Val219Leu) or likely pathogenic (n=4, specifically: *MYBPC3* p.Asn1257Lys; *MYH7* p.Arg1712Gln (in two unrelated individuals); *MYH7* p.Met877Ile) variants. An additional two individuals demonstrated pathogenic (*FLNC* c.6004+2T>C) and likely pathogenic (*FHL1* p.Cys221Tyr) variants in genes beyond those considered to be the core sarcomere genes. The emergence of pathogenic/likely pathogenic variants in the BRRD cohort demonstrates the evolving and dynamic nature of variant classification.

6.72% [95% CI:4.80-9.34%] of individuals recruited to GeL with HCM carried either a pathogenic (n=17) or likely pathogenic variant (n=15).

3.2.3 Genes implicated in HCM

The established finding that *MYH7* and *MYBPC3* yield the largest proportion of pathogenic/likely pathogenic variants associated with HCM is further replicated in the OMGL-HCMR cohort. In the OMGL-HCMR data, almost half (49.7% [95% CI: 47.1 – 52.3%]) of all likely pathogenic/pathogenic variants are attributable to pathogenic *MYBPC3* variants, however there is evidence of high allelic heterogeneity, with most likely pathogenic/pathogenic variants only observed once (Figure 3.1).

3.2.4 Sarcomere negative

Registry data from the Sarcomere Human Cardiomyopathy Registry (SHaRe) indicates that the overall composite for adverse outcomes (including stroke, heart failure, arrhythmia, transplant and others) for individuals diagnosed with HCM but in whom a pathogenic, likely pathogenic or variant of uncertain significance is not detectable are more favourable than for pathogenic/likely pathogenic variant carriers [HR:0.51 [95% CI:0.44-0.58]; p-value <0.001].[56] No pathogenic, likely pathogenic or variant of uncertain significance were detectable in 60.7% [95% CI: 59.4-62.0%] of the OMGL-HCMR cohort and 84.2% [95% CI: 78.8 – 88.6%] of the BRRD cohort (Figure 3.2).

3.2.5 Multiple variants

Individuals with HCM who carry multiple pathogenic/likely pathogenic variants are reported to experience a higher risk of adverse outcomes, including ventricular arrhythmia [HR: 4.15 [95% CI: 1.96-8.76], p<0.001], heart failure [HR: 2.17 [95% CI: 1.23-3.72], p<0.01] and atrial fibrillation [HR: 1.90 [95% CI: 1.01-3.58], p<0.05], relative to gene-negative HCM.[56] When compared to HCM cases carrying one pathogenic/likely pathogenic variant, individuals with multiple pathogenic/likely pathogenic variants have an increased risk of stroke [HR:5.08 [95% CI: 2.05 – 12.63]]

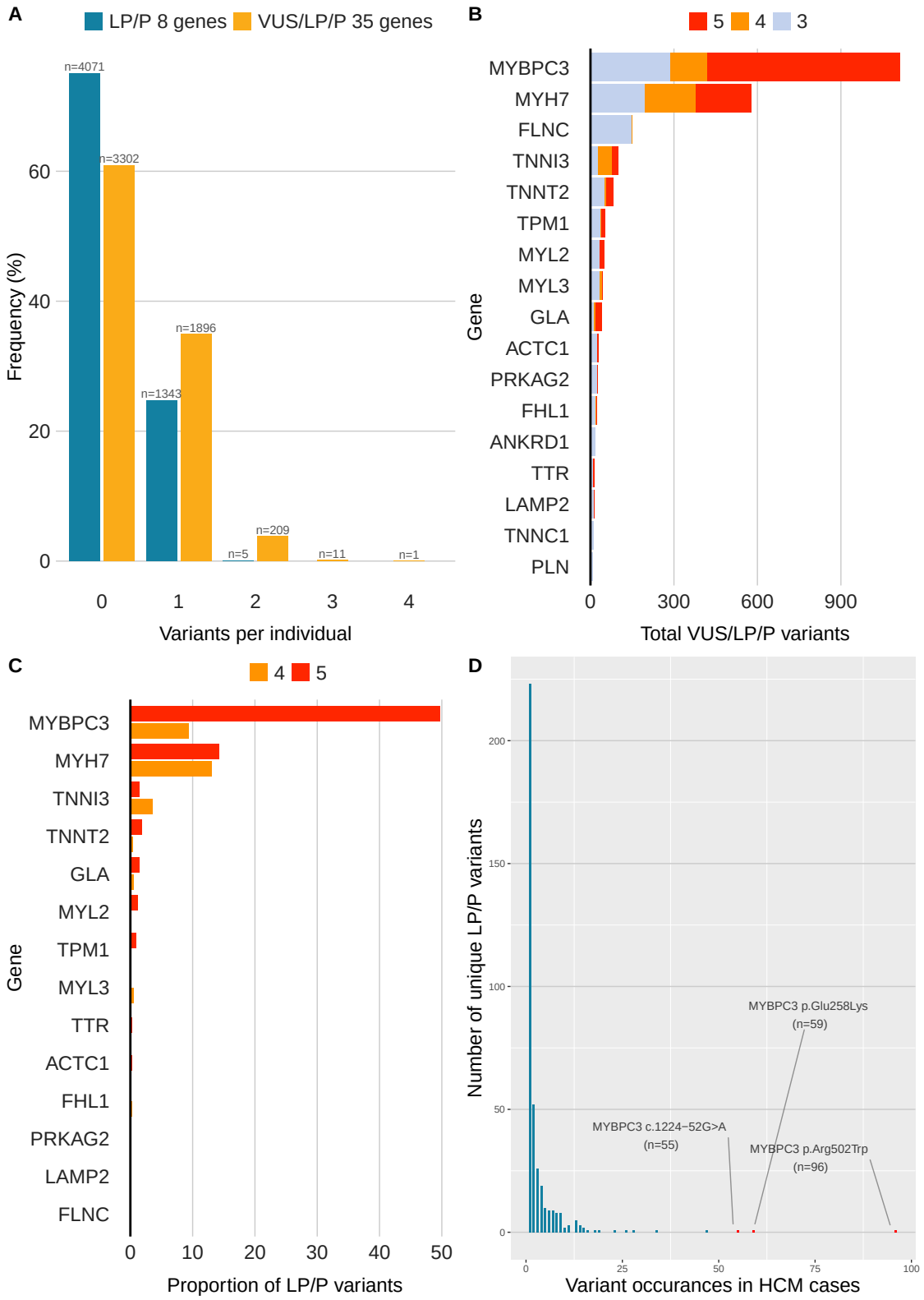


Figure 3.1: Evaluation of disease-causing variants in the OMGL-HCMR cohort. Panel A: Distribution of number of disease-causing variants per individual. Panel B: Genes harbouring variants of uncertain significant (3), likely pathogenic (4), or pathogenic (5) variants. Panel C: Proportion of likely pathogenic (4) and pathogenic (5) variants per gene. Panel D: Histogram demonstrating that the majority of likely pathogenic and pathogenic variants are only observed once. Frequently observed variants are highlighted. Abbreviations: LP: likely pathogenic; P: pathogenic; VUS: variant of uncertain significance

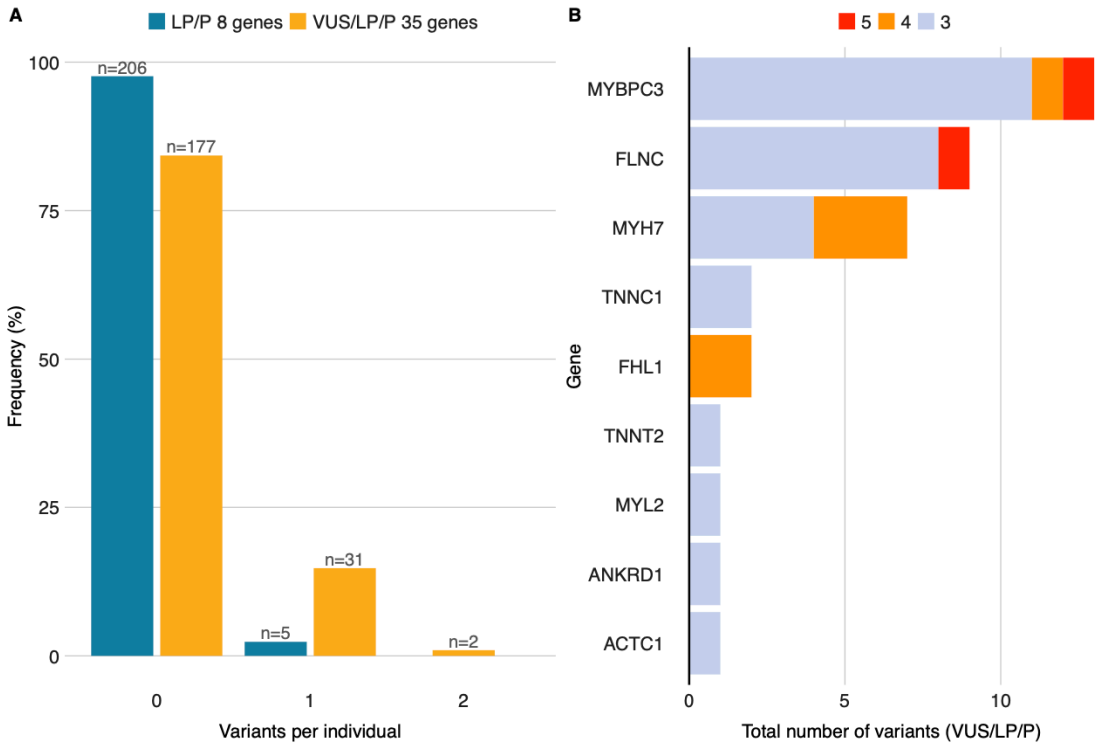


Figure 3.2: Evaluation of variants of disease causing variants in BRRD Panel A: Distribution of number of disease-causing variants per individual. Panel B: Genes harbouring variants of uncertain significant (3), likely pathogenic (4), or pathogenic (5) variants.Abbreviations: LP: likely pathogenic; P: pathogenic; VUS: variant of uncertain significance

and cardiac transplantation or left ventricular assist device [HR:7.48 [95% CI: 2.73-20.47]].[56] Prior to the adoption of contemporary variant classification methods, there was an assumption that ∼8% of individuals carrying a pathogenic/likely pathogenic variant (gene-positive) also carried a second pathogenic/likely pathogenic variant.[176] Re-evaluation of these data, with application of contemporary variant classification methods, revised the prevalence of double mutation carriers down to ∼0.4% of gene-positive individuals.[176] Data from the OMGL-HCMR cohort replicate these estimates (literature: 3/2494 vs. OMGL-HCMR:5/5,393; p-value = 1), and suggests only ∼1:1000 HCM cases (0.092% [95% CI: 0.039 – 0.22%]) carry more than one pathogenic/likely pathogenic variant (Table 3.4).

Gene	Transcript	HGVS.c	HGVS.p	ACMG	AF in gnomAD	Consequence	Count
Core sarcomeric genes (*MYBPC3, MYH7, TNNI3, TNNT2, MYL2, MYL3, ACTC1, TPM1*)							
MYBPC3	NM_000256.3	c.655G>C	p.Val219Leu	5	NA	Missense variant & splice region variant	1
MYH7	NM_000257.2	c.5135G>A	p.Arg1712Gln	4	1.77E-05	Missense variant	2
MYBPC3	NM_000256.3	c.3771C>A	p.Asn1257Lys	4	NA	Missense variant	1
MYH7	NM_000257.2	c.2631G>C	p.Met877Ile	4	NA	Missense variant	1
MYBPC3	NM_000256.3	c.2429G>A	p.Arg810His	3	5.70E-05	Missense variant	2
MYBPC3	NM_000256.3	c.3815-10T>G	-	3	4.22E-06	Intron variant	2
ACTC1	NM_005159.4	c.435T>A	p.Tyr145Ter	3	NA	Stop gained	1
MYBPC3	NM_000256.3	c.2441_2443del	p.Lys814del	3	4.82E-05	Inframe deletion	1
MYBPC3	NM_000256.3	c.2618C>A	p.Pro873His	3	6.80E-05	Missense variant	1
MYBPC3	NM_000256.3	c.3005G>A	p.Arg1002Gln	3	5.24E-05	Missense variant	1
MYBPC3	NM_000256.3	c.3470C>T	p.Pro1157Leu	3	3.65E-05	Missense variant	1
MYBPC3	NM_000256.3	c.3739G>A	p.Asp1247Asn	3	NA	Missense variant	1
MYBPC3	NM_000256.3	c.3798C>G	p.Cys1266Trp	3	NA	Missense variant	1
MYBPC3	NM_000256.3	c.909-7G>A	-	3	NA	Splice region variant & intron variant	1
MYH7	NM_000257.2	c.1405G>A	p.Asp469Asn	3	1.26E-05	Missense variant & splice region variant	1
MYH7	NM_000257.2	c.2273T>C	p.Phe758Ser	3	NA	Missense variant	1
MYH7	NM_000257.2	c.5342G>A	p.Arg1781His	3	1.76E-05	Missense variant	1
MYH7	NM_000257.2	c.964T>A	p.Ser322Thr	3	NA	Missense variant	1
MYL2	NM_000432.3	c.141C>A	p.Asn47Lys	3	2.09E-04	Missense variant	1
TNNT2	NM_001276345.1	c.815A>G	p.Asn272Ser	3	5.94E-06	Missense variant	1
Additional genes included on 35 gene cardiomyopathy panel							
FLNC	NM_001458.4	c.6004+2T>C	-	5	NA	Splice donor variant	1
FHL1	NM_001159702.2	c.662G>A	p.Cys221Tyr	4	NA	Missense variant	1
ANKRD1	NM_014391.2	c.368C>T	p.Thr123Met	3	3.24E-04	Missense variant	1
FLNC	NM_001458.4	c.2650G>T	p.Val884Phe	3	8.16E-06	Missense variant	1
FLNC	NM_001458.4	c.34C>G	p.Leu12Val	3	NA	Missense variant	1
FLNC	NM_001458.4	c.3650_3652del	p.Ser1217_Pro1218delinsThr	3	NA	Inframe deletion	1
FLNC	NM_001458.4	c.4061G>A	p.Arg1354Gln	3	1.92E-05	Missense variant	1
FLNC	NM_001458.4	c.4277G>A	p.Arg1426Gln	3	1.24E-05	Missense variant	1
FLNC	NM_001458.4	c.4504A>G	p.Thr1502Ala	3	NA	Missense variant	1
FLNC	NM_001458.4	c.6310G>A	p.Glu2104Lys	3	4.12E-06	Missense variant	1
FLNC	NM_001458.4	c.7205C>A	p.Ala2402Asp	3	NA	Missense variant	1
TNNC1	NM_003280.2	c.210C>T	p.Gly70%3D	3	1.18E-04	Synonymous variant	1
TNNC1	NM_003280.2	c.23C>T	p.Ala8Val	3	1.52E-05	Missense variant & splice region variant	1

Table 3.3: List of variants detected in BRRD Variants with an ACMG classification of either variant of uncertain significance, likely pathogenic or pathogenic. AF: allele frequency.

ID	Gene	c.HGVS	p.HGVS	ACMG
HCM_0142	*MYH7*	NM_000257.2:c.2631G>C	NP_000248.2:p.Met877Ile	4
	MYH7	NM_000257.2:c.1816G>A	NP_000248.2:p.Val606Met	5
HCM_0430	*MYL3*	NM_000258.2:c.517A>G	NP_000249.1:p.Met173Val	4
	MYL3	NM_000258.2:c.517A>G	NP_000249.1:p.Met173Val	4
HCR08783	*MYH7*	NM_000257.2:c.596C>T	NP_000248.2:p.Ala199Val	4
	TNNI3	NM_000363.4:c.434G>A	NP_000354.4:p.Arg145Gln	4
HCR13502	*TNNT2*	NM_001276345.1:c.890G>A	NP_001263274.1:p.Trp297Ter	5
	MYBPC3	NM_000256.3:c.551dup	NP_000247.2:p.Lys185GlufsTer56	5
HCR31823	*MYBPC3*	NM_000256.3:c.3233G>A	NP_000247.2:p.Trp1078Ter	5
	MYBPC3	NM_000256.3:c.3233G>A	NP_000247.2:p.Trp1078Ter	5

Table 3.4: Cases with multiple likely pathogenic or pathogenic variants

3.2.6 Frequently observed variants

Despite high levels of allelic heterogeneity, several variants are seen relatively frequently in HCM, attributable to both founder effects from a common ancestor and recurrent variants (Table 3.5).[177] Variants previously reported as founder mutations detected within this cohort, include: *MYBPC3* p.Trp792ValfsTer41 (Netherlands)[178], and *MYBPC3* p.Pro955ArgfsTer95141 (Netherlands) and *MYH7* p.Ala797Thr (South Africa).

The most frequently observed pathogenic variant is *MYBPC3* p.Arg502Trp, accounting for 1.82% [95% CI: 1.49 – 2.21%] of observed allelic heterogeneity (n=96/5,285) in OMGL-HCMR, and replicating prior literature findings.[8, 55, 179]

MYBPC3 c.1224-52G>A

A relatively novel finding relates to *MYBPC3* c.1224-52G>A, the third most frequently observed pathogenic variant in OMGL-HCMR, which accounts for 1.04% [95% CI: 0.80 – 1.35%] of HCM. Given the intronic location of *MYBPC3* c.1224-52G>A, most gene panel sequencing capture methods fail to detect *MYBPC3* c.1224-52G>A. This finding was only possible because the OMGL applied a customised capture method that extended into this deep intronic space, based on a prior hypothesis that splice-site disruption may cause HCM in genes where loss-of-function is a recognised mechanism, such as in *MYBPC3*. *In silico* splice site tools (specifically SpliceSiteFinder-like, MaxEntScan, NNSplice and Human Splicing Finder), suggested *MYBPC3* c.1224-52G>A would result in the formation of a cryptic splice acceptor site 50 nucleotides upstream of the native site.[180] These *in silico* predictions were confirmed when Sanger sequencing was performed on an aberrant gene product (Figure 3.3). Specifically, OMGL employees performed gel fractionation on cDNA that had been reverse transcribed from lymphocyte-derived RNA from 2 affected individuals harbouring the *MYBPC3* c.1224-52G>A. Polymerase chain reaction used primers directed towards exon 12 and exon 16. Electrophoresis, performed on an agarose gel, revealed two gene products: a normal fragment (323 bp) and an aberrant fragment (375 bp). These two PCR amplicons

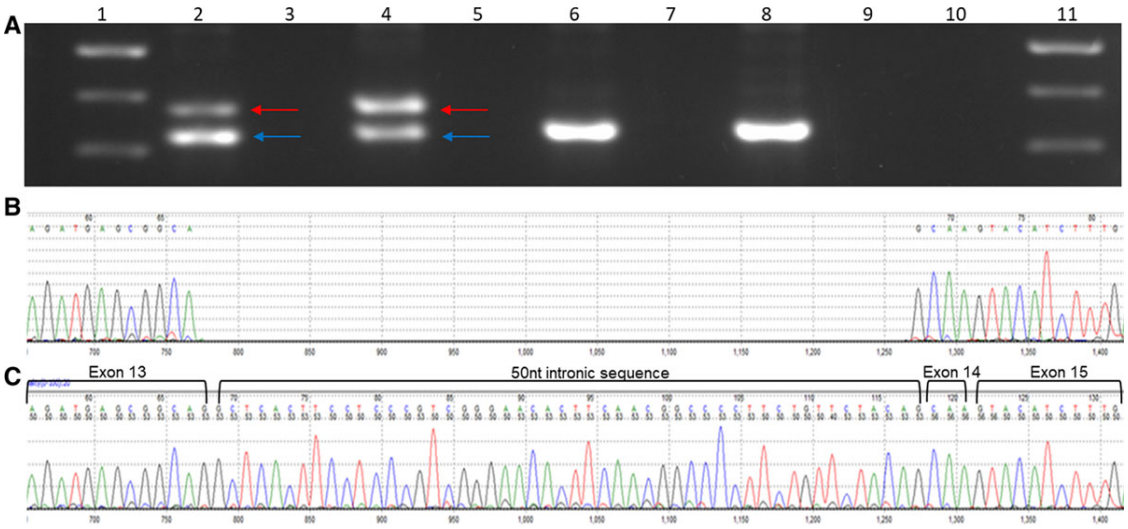


Figure 3.3: *MYBPC3* c.1224-52G>A RNA studies Panel A: PCR amplicons derived from cDNA, generated from the RNA of 2 affected individuals heterozygous for the *MYBPC3* c.1224-52A>G (lanes 2 and 4, and corresponding negative controls in lanes 3 and 5). Positive controls in lanes 6 and 8, with corresponding negative controls in lanes 7 and 9. Lane 1 andd 11 contain a 100 base pair ladder (500 bp [dense band], 400 bp, and 300 bp bands shown). Blue arrow corresponds with normal fragment (323 bp), as seen in controls, and the red arrow corresponds to the aberrant fragment (375 bp). Panels B and C: Sanger sequencing results from the wild-type (B) and aberrant polymerase chain reaction product derived from cDNA of an affected individual harbouring MYBPC3 c.1224-52A>G (C). Panel C demonstrates a 50-nucleotide intronic inclusion and confirms *in silico* splice site predictions. Figure reproduced from original publication, descending from this thesis, that describes the work [180]

were then gel-purified and Sanger sequenced. This revealed a 50-nucleotide intronic inclusion from intron 13 between exon 13 and exon 14 in the messenger RNA, and led to a frameshift in the amino acid sequence by introducing a premature termination codon at position 438 (p.Ser408fs*31). [180]

Given that most clinical diagnostic gene panel sequencing methodologies do not capture intronic regions, and the original description of *MYBPC3* c.1224-52G>A was derived from a small, family-based genome sequencing project, the discovery that *MYBPC3* c.1224-52G>A is a relatively prevalent cause of HCM is of importance. This finding should prompt healthcare systems to re-evaluate their current screening practices.[74]

GENE	c.HGVS	p.HGVS	ACMG	OMGL (n=2,758)		HCMR (n=2,527)		Total (n=5,285)	
				Count	%	Count	%	Count	%
MYBPC3	c.1504C>T	p.Arg502Trp	5	58	2.10 (1.63 - 2.71)	38	1.50 (1.10 - 2.06)	96	1.82 (1.49 - 2.21)
MYBPC3	c.772G>A	p.Glu258Lys	5	23	0.83 (0.56 - 1.25)	36	1.42 (1.03 - 1.97)	59	1.12 (0.87 - 1.44)
MYBPC3	c.1224-52G>A	-	5	32	1.16 (0.82 - 1.63)	23	0.91 (0.61 - 1.36)	55	1.04 (0.80 - 1.35)
MYBPC3	c.2373dup	p.Trp792ValfsTer41	5	11	0.40 (0.22 - 0.71)	37	1.46 (1.06 - 2.01)	48	0.91 (0.69 - 1.20)
MYBPC3	c.1624+4A>T	-	5	19	0.69 (0.44 - 1.07)	15	0.59 (0.36 - 0.98)	34	0.64 (0.46 - 0.90)
MYBPC3	c.1624G>C	p.Glu542Gln	5	15	0.54 (0.33 - 0.90)	13	0.51 (0.30 - 0.88)	28	0.53 (0.37 - 0.76)
MYH7	c.2389G>A	p.Ala797Thr	5	12	0.44 (0.25 - 0.76)	14	0.55 (0.33 - 0.93)	26	0.49 (0.34 - 0.72)
MYBPC3	c.1484G>A	p.Arg495Gln	4	8	0.29 (0.15 - 0.57)	15	0.59 (0.36 - 0.98)	23	0.44 (0.29 - 0.65)
MYBPC3	c.1928-2A>G	-	5	7	0.25 (0.12 - 0.52)	12	0.47 (0.27 - 0.83)	19	0.36 (0.23 - 0.56)
MYH7	c.1988G>A	p.Arg663His	5	7	0.25 (0.12 - 0.52)	11	0.44 (0.24 - 0.78)	18	0.34 (0.22 - 0.54)
MYH7	c.5135G>A	p.Arg1712Gln	4	5	0.18 (0.08 - 0.42)	11	0.44 (0.24 - 0.78)	16	0.30 (0.19 - 0.49)
MYBPC3	c.655G>C	p.Val219Leu	5	10	0.36 (0.20 - 0.67)	5	0.20 (0.08 - 0.46)	15	0.28 (0.17 - 0.47)
MYBPC3	c.927-9G>A	-	5	3	0.11 (0.04 - 0.32)	12	0.47 (0.27 - 0.83)	15	0.28 (0.17 - 0.47)
GLA	c.644A>G	p.Asn215Ser	5	10	0.36 (0.20 - 0.67)	4	0.16 (0.06 - 0.41)	14	0.26 (0.16 - 0.44)
MYBPC3	c.3330+5G>C	-	5	8	0.29 (0.15 - 0.57)	6	0.24 (0.11 - 0.52)	14	0.26 (0.16 - 0.44)
TNNI3	c.433C>T	p.Arg145Trp	5	9	0.33 (0.17 - 0.62)	5	0.20 (0.08 - 0.46)	14	0.26 (0.16 - 0.44)
MYBPC3	c.1227-13G>A	-	5	6	0.22 (0.10 - 0.47)	7	0.28 (0.13 - 0.57)	13	0.25 (0.14 - 0.42)
MYBPC3	c.2096del	p.Pro699GlnfsTer55	5	6	0.22 (0.10 - 0.47)	7	0.28 (0.13 - 0.57)	13	0.25 (0.14 - 0.42)
MYH7	c.2609G>A	p.Arg870His	5	7	0.25 (0.12 - 0.52)	6	0.24 (0.11 - 0.52)	13	0.25 (0.14 - 0.42)
MYH7	c.2681A>G	p.Glu894Gly	5	5	0.18 (0.08 - 0.42)	8	0.32 (0.16 - 0.62)	13	0.25 (0.14 - 0.42)
MYBPC3	c.1224-19G>A	-	4	6	0.22 (0.10 - 0.47)	5	0.20 (0.08 - 0.46)	11	0.21 (0.12 - 0.37)
MYBPC3	c.2308G>A	p.Asp770Asn	4	4	0.15 (0.06 - 0.37)	7	0.28 (0.13 - 0.57)	11	0.21 (0.12 - 0.37)
MYBPC3	c.2864_2865del	p.Pro955ArgfsTer95	5	7	0.25 (0.12 - 0.52)	4	0.16 (0.06 - 0.41)	11	0.21 (0.12 - 0.37)
MYH7	c.1063G>A	p.Ala355Thr	4	6	0.22 (0.10 - 0.47)	5	0.20 (0.08 - 0.46)	11	0.21 (0.12 - 0.37)
MYBPC3	c.177_187del	p.Glu60AlafsTer49	5	5	0.18 (0.08 - 0.42)	5	0.20 (0.08 - 0.46)	10	0.19 (0.10 - 0.35)
TNNI3	c.422G>A	p.Arg141Gln	4	6	0.22 (0.10 - 0.47)	4	0.16 (0.06 - 0.41)	10	0.19 (0.10 - 0.35)

Table 3.5: Most frequently observed likely pathogenic or pathogenic variants across the OMGL and HCMR cohorts

3.2.7 Haplotype analysis of frequently observed variants

To establish whether a frequently observed variant is attributable to founder or recurrent effects requires haplotype analysis. Founder variants have a specific set of properties that differentiate them from recurrent variants. By definition, founder variants have been inherited from a common ancestor and individuals possessing a founder variant will share, not only the founder variant, but also the abutting nucleotide sequences. Therefore, the segment of DNA upon which the founder variant is present is considered identical by descent (IBD), presuming there is no evidence of recombination. Determining whether a variant truly demonstrates founder effects is reliant on access to accompanying genealogical information. In HCM, several founder variants have previously been described including *MYBPC3* c.2373dup p.Trp792ValfsTer41, which accounts for ∼25% of HCM detected in the Netherlands.[181] Individuals possessing a HCM founder variant tend to develop milder forms of disease, often slightly later in life than highly penetrant variants, and with no impact on fecundity the founder variant is vertically transmitted.

Recurrent variants are suspected when nucleotide sequences neighbouring a frequently observed variant differs between individuals, and may therefore be attributable to repeated *de novo* events in the context of a mutational hotspot, classically, CpG dinucleotides subject to methylation.

For variants observed on more than ten occasions in the HCMR cohort, haplotypes were constructed to evaluate for evidence of founder effects. For each rare variant, analysis was restricted to only include carriers of the given rare variant. Haplotypes were generated by merging rare variant data, derived from gene panel sequencing, with genotypes, derived from the Affymetrix Precision Medicine genotyping array, located in the surrounding ±100 kb region. Haploview (version 1.0) was used to evaluate haplotype structure in rare variant carriers.[182] Markers were selected when the: Hardy-Weinberg p-value threshold <0.001, minimum genotyping rate was > 75% and minor allele frequency >0.1.

Haplotypes were constructed for individuals possessing a frequently observed pathogenic/likely pathogenic variant (defined here as an allele count exceeding ten

in the HCMR cohort) using and common variants (MAF > 0.1) captured by a genotyping array in the surrounding ±100 kilobase region (Figures 3.4, 3.5 and 3.6). Whilst the haplotype structure can be graphically represented using HaploView, the haplotype structure is dependent on which SNPs are selected for inclusion and the portion of flanking sequence included. To further advance the evaluation of these haplotypes would require additional mathematical consideration using coalescent theory, allowing for recombination.[183] Nevertheless, here I report in detail the visual examination of the haplotype structure surrounding *MYBPC3* p.Arg502Trp.

The background haplotype patterns observed for *MYBPC3* p.Arg502Trp are generated from common (MAF > 0.1) SNPs that have likely arisen prior to the human expansion out of Africa, and as such, are now globally distributed with tens of thousands of generations of mutations and recombinations explaining their contemporary haplotype patterns. Prior studies designed to evaluate for the presence of founder variants have benefited from studying individuals, in whom a rare pathogenic variant is detected, from within an isolated population of relatively narrow diversity.[178, 184] In contrast, the HCMR cohort were recruited from across two continents (America and Europe), with inclusion of individuals from diverse ancestral backgrounds, which collectively makes inferences regarding the presence of founder variants challenging through application of a cladistic model. Whilst it is possible to attempt to describe each haplotype using a cladistic model, it is unlikely to reveal the full extent of a pathogenic variant's diversity. Nevertheless, analysis of the 38 individuals carrying *MYBPC3* p.Arg502Trp suggests several distinct clades. The most frequent clade encompasses four distinct haplotypes: H2 (28.9% [11/38]), H5 (10.5% [4/38]), H11(2.6% [1/38]) and H8 (5.0% [2/38]) (Figure 3.4), and accounts for 47.4% [95% CI: 32.5 – 62.7] of *MYBPC3* p.Arg502Trp carriers (n=18/38). It appears H2 may have arisen from haplotypes H5 and H11, and led to the formation of H8. Given the frequency and distinct pattern of the haplotypes within this clade, relative to the most commonly observed null haplotype (H1), raises the possibility that haplotypes contributing towards this clade are descended from a common founder. Correlation with genealogical details would be necessary

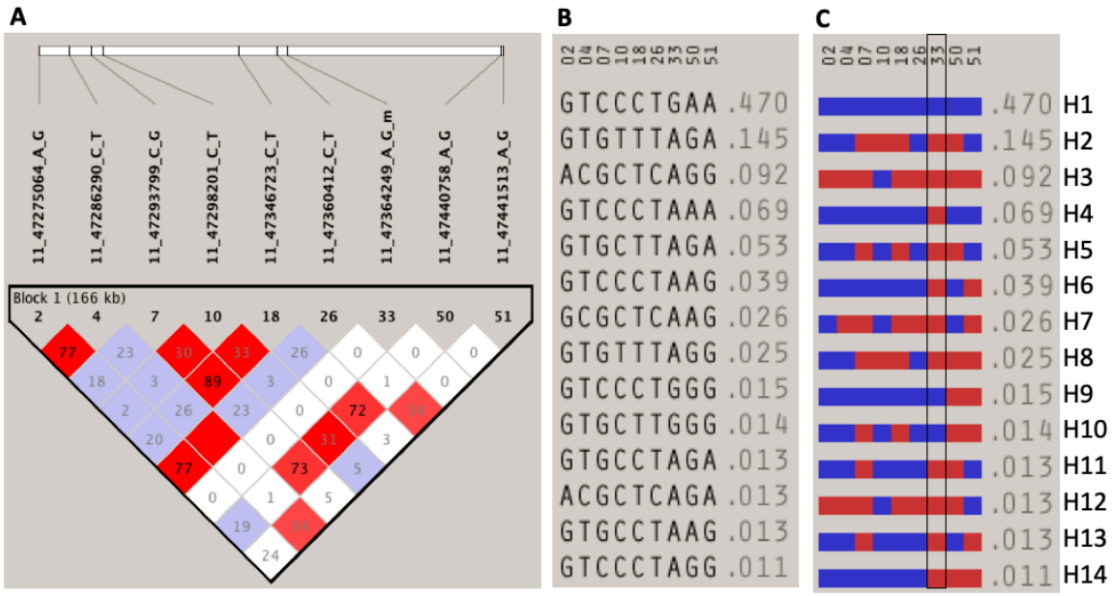


Figure 3.4: Haplotype analysis for NM_000256.3(*MYBPC3*):c.1504C>T (p.Arg502Trp) *MYBPC3* p.Arg502Trp is denoted here as 11-47364249-G-A_m and corresponds to number 33 in panels A, B and C. Wild type allele is G, alternate allele is A. Panel A illustrates correlation between genotyped SNPs ± 100kb *MYBPC3* p.Arg502Trp numerically using r^2, and visually using D prime (red indicates high D prime, white indicates low D prime, purple indicates high D prime and low LOD). Panel B and panel C represent the same information. Panel B is denoted using reference and alternative alleles, whilst panel C represents this visually (blue=reference allele and red=alternate allele). The black box in panel C identifies the marker for *MYBPC3* p.Arg502Trp. Labelling H1-H14 of distinct haplotypes.

to confirm this hypothesis. It is possible that *MYBPC3* p.Arg502Trp is also attributable to recurrent variants, with haplotype H4 appearing on the predominant null haplotype (H1) and H14 appearing on a less prevalent null haplotype (H9). Collectively, the H4 and H14 haplotypes account for 15.8% [95% CI: 7.4 – 30.4%] of *MYBPC3* p.Arg502Trp carriers in the HCMR cohort. Additional possible clades include: H3-H7-H12 (21.1% [95% CI: 11.1-36.3], n=8/38) and H6-H13 (10.5% [95% CI: 4.2 – 24.1], n=4/38). It is possible that haplotypes H6 and H13 arose from haplotype H4, which is speculated to be a recurrent variant. Haplotypes for other frequently observed pathogenic variants are presented in Figures 3.5 and 3.6, but analysis of these is haplotypes is not reported given the limitations described above.

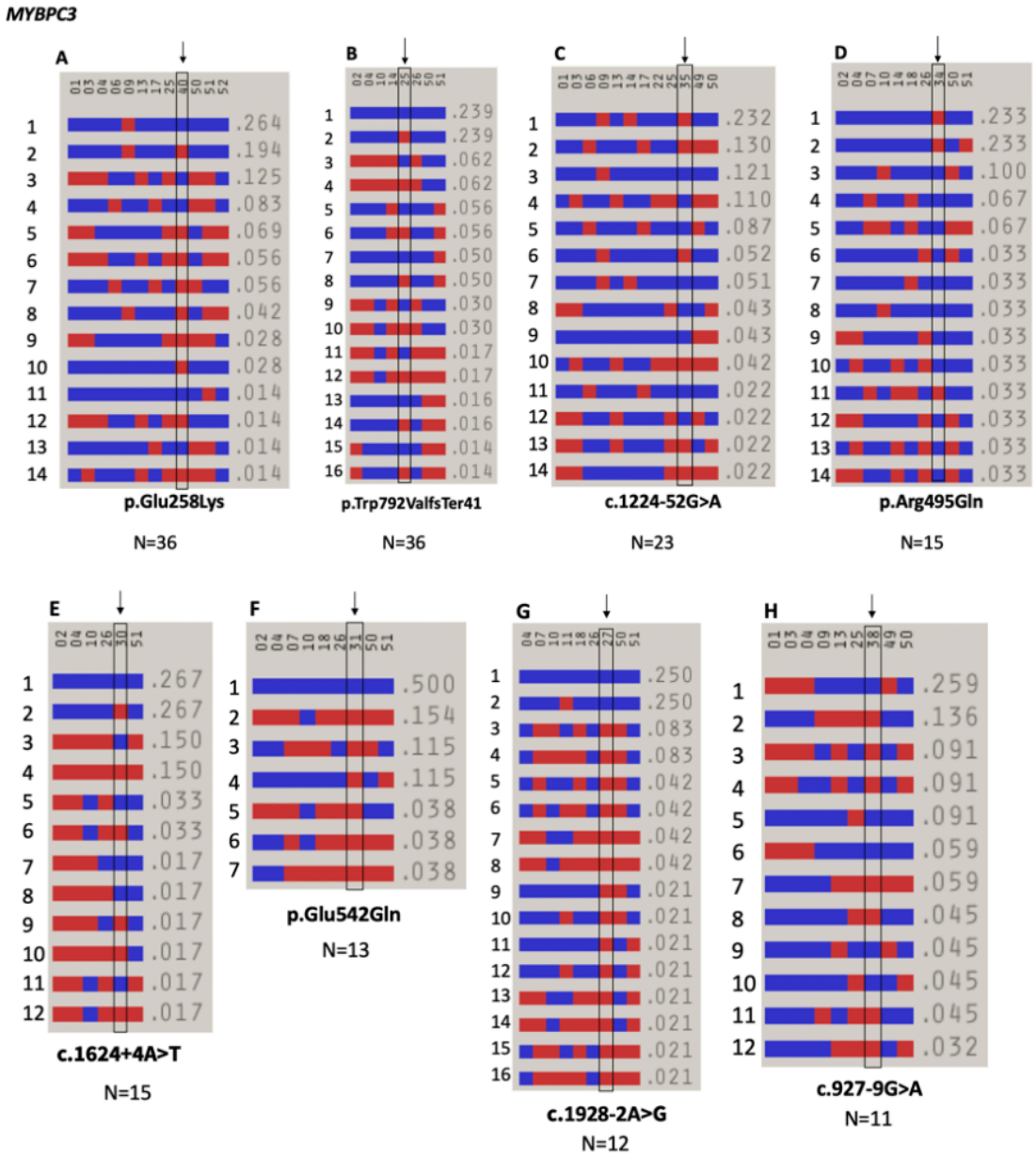


Figure 3.5: Haplotypes across frequently detected (n>10) pathogenic and likely pathogenic variants in *MYBPC3* from the HCMR cohort

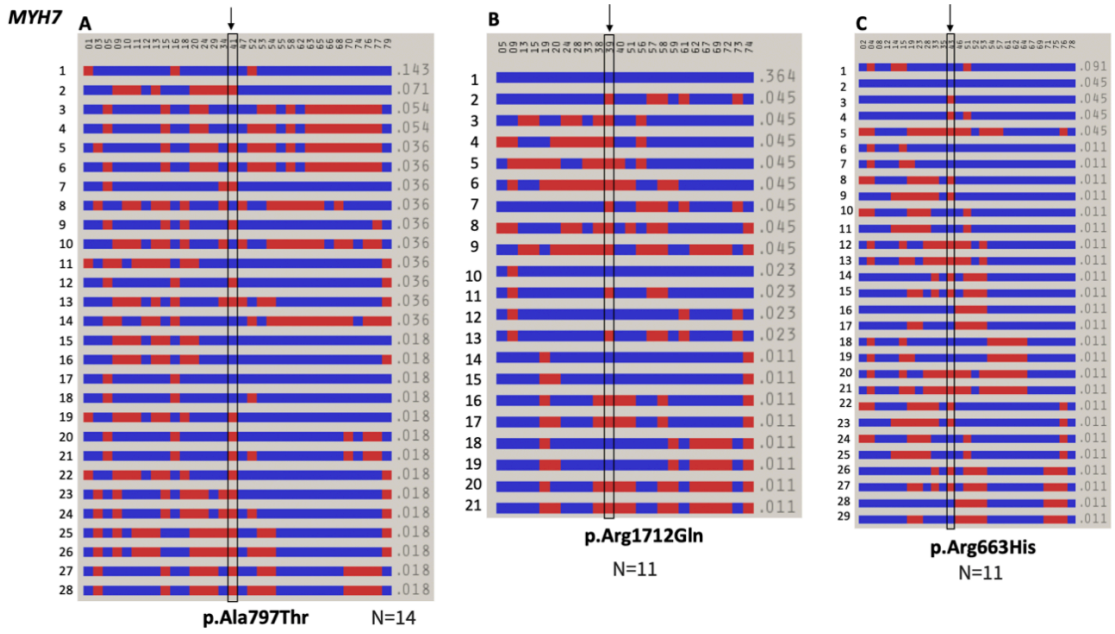


Figure 3.6: Haplotypes across frequently detected (n>10) pathogenic and likely pathogenic variants in *MYH7* from the HCMR cohort

3.3 Rare variant burden analysis

3.3.1 Burden testing

It is plausible that genes harbouring disease-causing variants exist beyond those that have already been established but remain undetected. To address this hypothesis, burden testing can be performed across genes assumed to have a causal role in cardiomyopathy. Analysis of 51 genes, determined to be implicated in HCM by a committee informing the development of Genomics England's PanelApp for HCM (https://panelapp.genomicsengland.co.uk/), was performed in the BRRD cohort (240 sarcomere negative HCM cases vs. 6,229 controls).[71, 121] Analysis performed by Thomson et al (2019) suggested these additional genes were not a major unaccounted cause for monogenic HCM. However, this analysis was only adequately powered (86% powered) to detect an odds ratio of 10 between cases and controls, at an α threshold of 0.001. It is therefore reasonable to further evaluate the 35 cardiomyopathy-associated genes, for which individual level gene panel sequence data is available; the power to detect more subtle effects will be

improved, given the increased case cohort size.

3.3.2 Methodological considerations

Burden testing is a statistical methodology that compares the aggregated total of pre-defined genetic variants within a given region of the genome, between cases and controls.[97] A range of software packages are available to perform burden testing, but these tend to require genome-wide data, and employ regression-based methods that correct for covariates. Here, gene-based burden testing was performed using gene-panel data to quantify associations between HCM cases and controls, stratified by variant consequence, for rare genetic variants with a filtering allele frequency < 0.0001 across each discrete population within gnomAD (POPMAX) and across the entire non-overlapping gnomAD and TOPMED population.

Fisher's exact test was selected as an appropriate analysis method given that the marginal counts derived from rare variant burden testing are often low (i.e. less than 5 counts), and often zero. However, it could be argued that Fisher's exact test is an overly conservative method to evaluate rare variant count data, given that the 2×2 contingency table, used to compute the conditional probability of observing the stated marginal counts, is doubly conditioned by both variant carrier status and affected status. In most experimental designs the nominal variables are either unconditioned or singly conditioned. Here, p-values derived from Fisher's exact test represent the proportion of permutation outcomes more extreme than the observed marginal counts. As the test is doubly constrained, rejecting the null hypothesis (i.e. p-value $< \alpha$ threshold) may be more challenging than an alternate "exact", permutation based, methodology. Exact permutation methods were traditionally considered to be computationally intensive, but advances in modern computing make this less of a concern, unless tens of thousands of cases and controls are to be analysed. Nevertheless, exact permutation methods are generally considered to provide small incremental gains and consequently, in these circumstances, Fisher's exact test was deemed appropriate.

Whilst rare variants have not previously demonstrated a protective effect for cardiomyopathy, given that I am testing a wider panel of cardiomyopathy genes, including those associated with dilated cardiomyopathy – a condition which demonstrates diametrically opposing molecular effects to HCM – it is plausible that a protective effect could be observed. As such, a two-sided test Fisher's exact test, with 95% confidence intervals, will be performed. To account for multiple testing, and reduce the chance of detecting a false-positive association (type I error), a Bonferroni correction was performed by dividing a predefined α of 0.05 by the number of tests performed (n=105). This resulted in a significance threshold of 4.76×10^{-4}.

3.3.3 Synonymous variant burden

The underlying ancestry of the OMGL cohort is unknown, but has been estimated to be ∼80% European using principal components from gene panel sequence data (Chapter 2). Without access to genome-wide genotyping, average ancestral differences cannot be accounted for and may lead to false-positive associations. However, a person's genome is a mosaic of ancestral chromosomes and it could be argued that rather than the average genetic ancestry, it is the local ancestry, specific to the HCM-related genes, that is of greatest importance.[185, 186] Nevertheless, to further evaluate the local genetic ancestry would also require additional genotyping of ancestrally informative intronic and extragenic variants in the genomic regions of interest, which is not currently feasible.

As such, alternative measures are required to assess for evidence of population stratification. It has previously been demonstrated that the rate of synonymous variants between cases and controls can be used as a reliable marker for inter-ancestry differences and considered analogous to a negative control group.[97, 187, 188]

To demonstrate that synonymous variants within the regions captured by the 35 gene cardiomyopathy panel are sufficient to discriminate between ancestral groups, analyses were performed across the MAF thresholds between 0.0001 and 0.1. I used individual-level data extracted from 1000 Genomes Phase 3 data, annotated using VEP, with inclusion of gnomAD allele frequencies. The synonymous burden

per individual was then evaluated across the OMGL, HCMR and T2DM cohorts to assess for evidence of population stratification.

As there were insufficient variants per-population group to facilitate meaningful analysis, it was not possible to discriminate between ancestral groups, due to low variant counts per population group, at a MAF threshold of 0.0001 (see Figure 3.7). At a MAF threshold of 0.1, differences between ancestral groups emerged (Kruskal-Wallis chi-squared = 229.38, 3 degrees of freedom, p-value $< 2.2\times10^{-16}$, given the data were not normally distributed (Levene's Test for Homogeneity of Variance p-value=2.2×10^{-7})).

No significant differences for the frequency of synonymous variants, with a MAF threshold <0.1, were observed between OMGL-$HCMR_{EU}$ cases and T2DM controls (non-parametric Wilcoxon rank sum test with continuity correction p-value = 0.9236, given that the data were not normally distributed: Levene's Test for Homogeneity of Variance p-value=2.07×10^{-3}). No extreme differences in common synonymous variant burden were detected between the OMGL, $HCMR_{EU}$ and T2DM cohorts (Figure 3.8).

3.3.4 Burden testing: OMGL vs $HCMR_{EU}$

Relative to the OMGL cohort (n= 2,757), $HCMR_{EU}$ (n= 1,980) appears enriched for variants known to cause HCM: an excess of both *MYBPC3* loss-of-function variants (OMGL: 6.45% vs. $HCMR_{EU}$: 12.9%; OR:0.50 [95% CI: 0.40 – 0.62]; p-value=1.54×10^{-10}) and *MYH7* missense variants (OMGL: 10.68% vs. $HCMR_{EU}$: 14.52%; OR:0.74 [95% CI: 0.61 – 0.89]; p- value=1.10×10^{-3}) were observed in the HCMR cohort, although only *MYBPC3* loss-of- function variants demonstrated statistical significance following Bonferroni correction (Figure 3.9). As anticipated, no systematic differences were observed between the OMGL and $HCMR_{EU}$ cohorts, with respect to the burden of missense or synonymous variants. This would indicate that the OMGL and HCMR cohorts are directly comparable. Differences in the rates of *MYBPC3* loss-of-function variants and *MYH7* missense variants may reflect a higher proportion of sarcomeric HCM being recruited into the HCMR

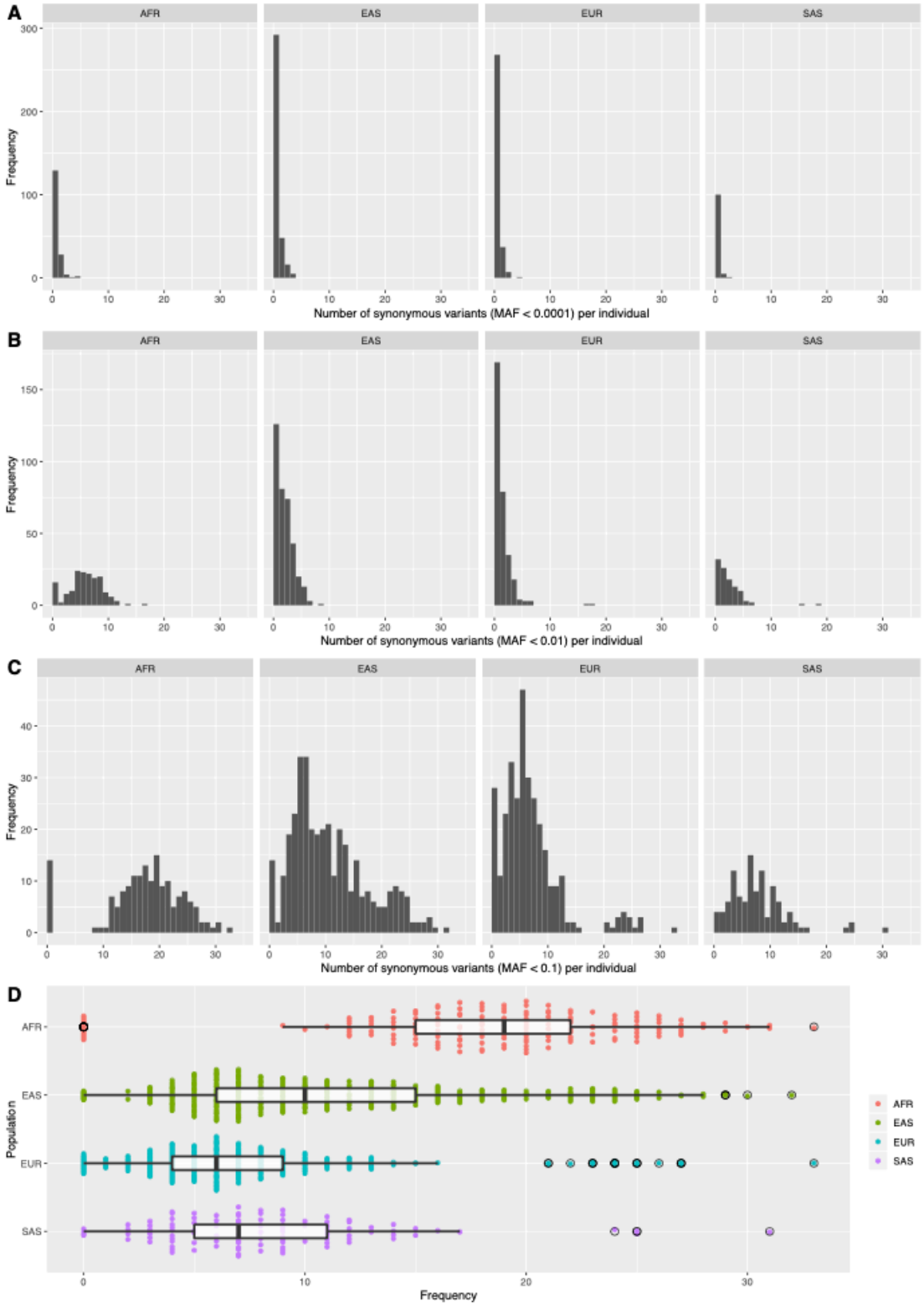


Figure 3.7: Distribution of synonymous variants per-individual across 35 cardiomyopathy related genes amongst unrelated individuals from the 1000 Genomes Phase 3 data Panel A: Synonymous variants with a minor allele frequency threshold < 0.0001; Panel B: Synonymous variants with a minor allele frequency < 0.01; Panel C: Synonymous variants with a minor allele frequency <0.1; Panel D: Ability to discriminate between ancestral groups using synonymous variants with a minor allele frequency threshold < 0.1. AFR: African; EAS: East Asian; EUR: European; SAS: South Asian.

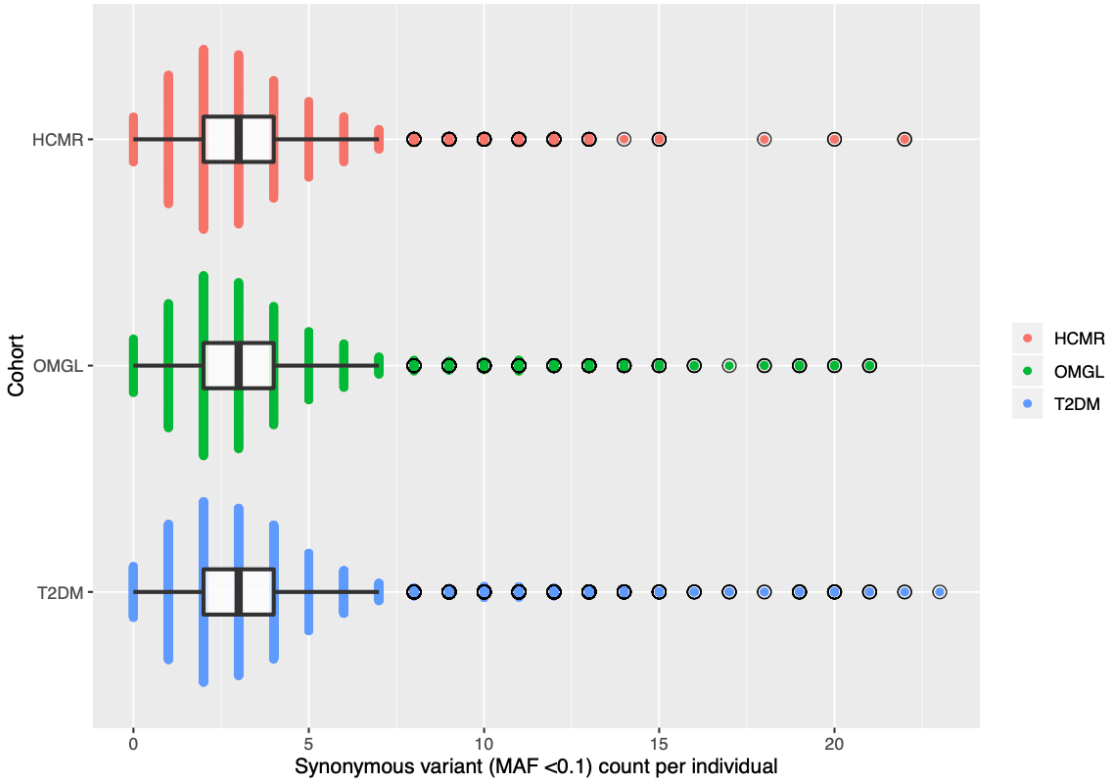


Figure 3.8: Comparison of synonymous variants, per individual, across cases (HCMR and OMGL) and controls (T2DM) indicates no extreme population stratification.

study. This might be anticipated, given that the OMGL samples are derived from a clinical diagnostic genetic testing centre that has fewer eligibility criteria than the HCMR study. The OMGL and HCMR_{EU} cohorts were merged to form the OMGL-HCMR_{EU} cohort (n=4,737).

3.3.5 Burden testing: OMGL- HCMR_{EU} vs T2DM

Alignment with prior observations

Using data derived from the OMGL-HCMR_{EU} cohort, 4,737 HCM cases were compared against 12,297 T2DM controls. This analysis involved the evaluation of individual-level sequence data from across 35 cardiomyopathy genes (∼211 kb) and improved on prior efforts that were reliant on clinical summary reports.[8] By processing the raw underlying data, quality control measures have been integrated, such as consideration for a negative control group (synonymous burden test), to help

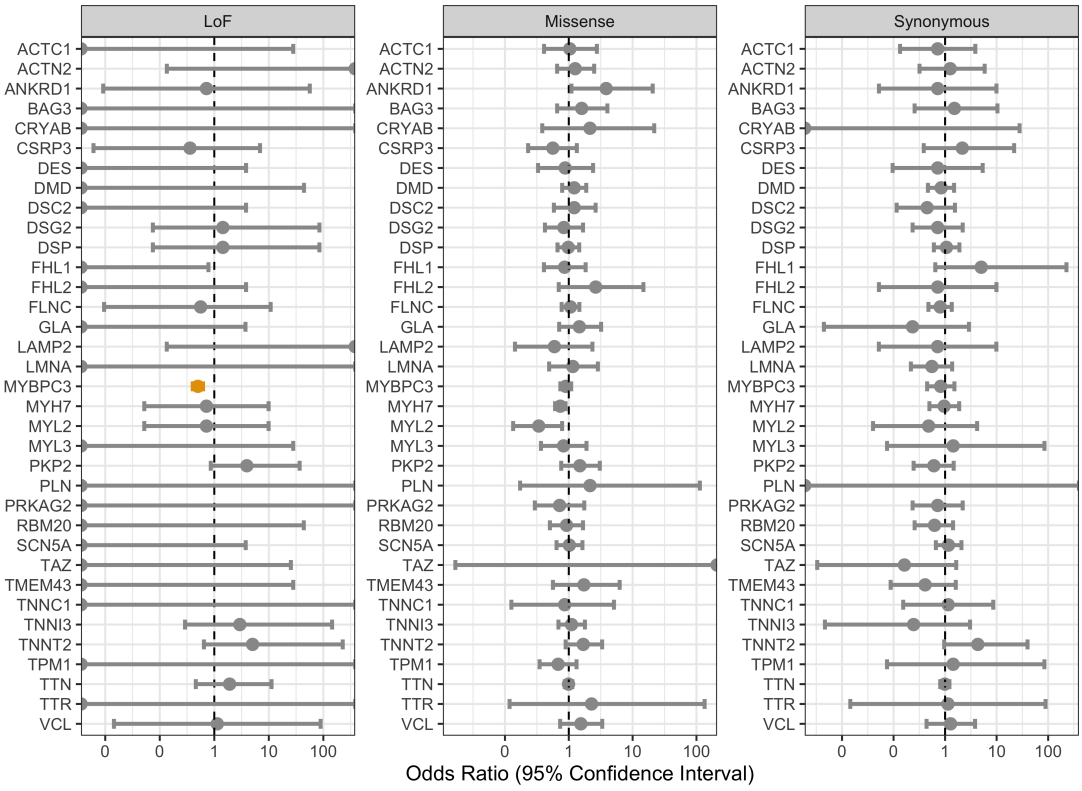


Figure 3.9: Results of burden testing in European HCM cases identified from the HCMR cohort compared against the OMGL HCM cohort Summary of burden testing across 35 cardiomyopathy associated genes. Analyses stratified by variant consequence into predicted loss-of- function variants (LoF), non-truncating variants, and synonymous variants. Data highlighted in yellow signifying significance following Bonferroni correction (α threshold = 4.76×10^{-4} (0.05/105 test performed))

facilitate the interpretation of these findings. The results presented here confirm many prior findings, including the observation that HCM cases were enriched for predicted loss of function variants in *MYBPC3*, and non-truncating variants in *ACTC1, FHL1, GLA, MYBPC3, MYH7, MYL2, MYL3, TNNI3, TNNT2* and *TPM1* (Figure 3.10 and Table 3.6).[8] Whilst it was anticipated that loss of function variants in *PLN* and *TNNT2*, and non-truncating variants in *PRKAG2* would demonstrate enrichment in HCM, these findings were not confirmed.[8] No *PLN* truncating variants were detected by LOFTEE in either cases or controls, and although *TNNT2* harboured an excess of truncating variants (HCM: 0.16% vs. T2DM: 0.04%; OR:4.16 [95% CI: 1.20–16.17]; p-value=1.12×10^{-2}), this was not statistically significant following Bonferroni correction. A lower proportion of

non-truncating *PRKAG2* variants detected in both HCM cases (1.01% vs. 0.51%; p-value=0.0095) and controls (0.52% vs. 0.33%; p-value=0.0088) resulted in less power to reject the null hypothesis for non-truncating *PRKAG2* variants [OR: 1.52 [95% CI: 0.88–2.58]; p-value=0.13] than prior analysis.[8] Overall, these results align with prior assumptions and indicate that the OMGL-HCMR cohort is suitable for further analyses investigating more complex genetic architecture models.

Novel findings

This analysis suggests HCM cases carry an excess of non-truncating variants across *CSRP3* (OR:3.10 [95% CI: 1.66 – 5.83]; p-value=1.84×10^{-4}), *FLNC* (OR: 1.39 [95% CI: 1.15 – 1.68]; p-value=4.61×10^{-4}), and high PSI *TTN* exons (OR:1.39 [95% CI: 1.28 – 1.51]; p-value=3.49×10^{-15}). Each of these findings is discussed in more detail below, to help contextualise the findings.

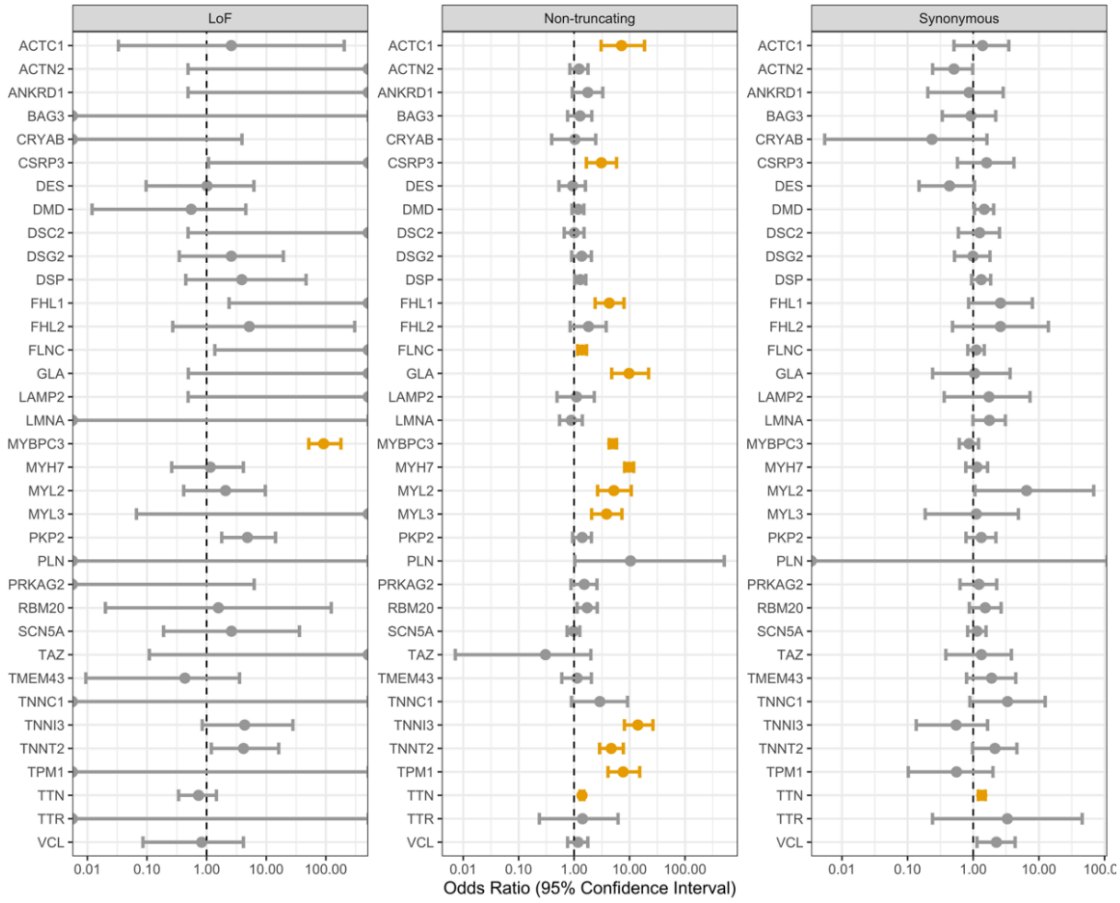


Figure 3.10: Results of burden testing in European HCM cases identified from the HCMR cohort compared against the T2DM cohort Individuals included in the HCMR cohort are not closely related and demonstrate European ancestry, as indicated by principal components analysis performed in relation to the 1000 Genomes ancestry. Controls derived from a T2DM consortia were not closely related and demonstrated European ancestry via principal components analysis. Summary of burden testing across 35 cardiomyopathy associated genes. Analyses stratified by variant consequence into predicted loss-of- function variants (LoF), non-truncating variants, and synonymous variants. Data highlighted in yellow signifying significance following Bonferroni correction (α threshold = 4.76×10^{-4} (0.05/105 test performed)).

	Predicted Loss of function (LOFTEE)				Non-truncating				Synonymous			
	Cases	Controls	OR (95% CI)	P-value	Cases	Controls	OR (95% CI)	P-value	Cases	Controls	OR (95% CI)	P-value
ACTC1	1/4737 (0.02%)	1/12297 (0.01%)	2.60 (0.03-203.49)	0.48	22/4737 (0.46%)	8/12297 (0.07%)	7.17 (3.07-18.62)	2.86E-07	8/4737 (0.17%)	15/12297 (0.12%)	1.39 (0.51-3.48)	0.49
ACTN2	2/4736 (0.04%)	0/12297 (0%)	-	-	44/4736 (0.93%)	93/12297 (0.76%)	1.23 (0.84-1.78)	0.25	11/4736 (0.23%)	56/12297 (0.46%)	0.51 (0.24-0.98)	0.040
ANKRD1	2/4734 (0.04%)	0/12297 (0%)	-	-	19/4734 (0.4%)	28/12297 (0.23%)	1.77 (0.93-3.28)	0.071	4/4734 (0.08%)	12/12297 (0.1%)	0.87 (0.2-2.86)	1.00
BAG3	0/3723 (0%)	0/12297 (0%)	-	-	24/3723 (0.64%)	62/12297 (0.5%)	1.28 (0.76-2.08)	0.31	7/3723 (0.19%)	25/12297 (0.2%)	0.92 (0.34-2.20)	1.00
CRYAB	0/4728 (0%)	4/12235 (0.03%)	0 (0-3.92)	0.58	8/4728 (0.17%)	20/12235 (0.16%)	1.04 (0.39-2.46)	1	1/4728 (0.02%)	11/12235 (0.09%)	0.24 (0.01-1.62)	0.20
CSRP3	3/4737 (0.06%)	0/12297 (0%)	-	-	25/4737 (0.53%)	21/12297 (0.17%)	3.10 (1.66-5.83)	0.00018	8/4737 (0.17%)	13/12297 (0.11%)	1.6 (0.57-4.16)	0.33
DES	2/4737 (0.04%)	5/12038 (0.04%)	1.02 (0.10-6.21)	1	20/4737 (0.42%)	54/12038 (0.45%)	0.94 (0.53-1.6)	0.90	6/4737 (0.13%)	35/12038 (0.29%)	0.43 (0.15-1.05)	0.056
DMD	1/3719 (0.03%)	6/12297 (0.05%)	0.55 (0.01-4.54)	1	93/3719 (2.5%)	262/12297 (2.13%)	1.18 (0.92-1.5)	0.18	54/3719 (1.45%)	122/12297 (0.99%)	1.47 (1.04-2.05)	0.024
DSC2	2/4731 (0.04%)	0/12296 (0%)	-	-	35/4731 (0.74%)	90/12296 (0.73%)	1.01 (0.66-1.51)	1.00	13/4731 (0.27%)	27/12296 (0.22%)	1.25 (0.59-2.51)	0.48
DSG2	3/4737 (0.06%)	3/12297 (0.02%)	2.6 (0.35-19.39)	0.36	39/4737 (0.82%)	74/12297 (0.6%)	1.37 (0.90-2.05)	0.11	16/4737 (0.34%)	42/12297 (0.34%)	0.99 (0.52-1.8)	1.00
DSP	3/4737 (0.06%)	2/12297 (0.02%)	3.90 (0.45-46.65)	0.14	116/4737 (2.45%)	234/12297 (1.9%)	1.29 (1.02-1.63)	0.026	57/4737 (1.2%)	112/12297 (0.91%)	1.33 (0.94-1.84)	0.085
FHL1	5/4734 (0.11%)	0/12297 (0%)	-	-	33/4734 (0.7%)	20/12297 (0.16%)	4.31 (2.40-7.93)	1.96E-07	8/4734 (0.17%)	8/12297 (0.07%)	2.60 (0.85-7.95)	0.088
FHL2	2/4737 (0.04%)	1/12297 (0.01%)	5.19 (0.27-305.84)	0.19	14/4737 (0.3%)	20/12297 (0.16%)	1.82 (0.85-3.79)	0.087	4/4737 (0.08%)	4/12297 (0.03%)	2.60 (0.48-13.95)	0.23
FLNC	3/3704 (0.08%)	0/12297 (0%)	-	-	171/3704 (4.62%)	413/12297 (3.36%)	1.39 (1.15-1.68)	0.00046	67/3704 (1.81%)	201/12297 (1.63%)	1.11 (0.83-1.47)	0.47
GLA	2/4686 (0.04%)	0/12297 (0%)	-	-	37/4686 (0.79%)	10/12297 (0.08%)	9.78 (4.76-22.06)	4.27E-13	4/4686 (0.09%)	10/12297 (0.08%)	1.05 (0.24-3.64)	1.00
LAMP2	2/4716 (0.04%)	0/12296 (0%)	-	-	11/4716 (0.23%)	26/12296 (0.21%)	1.10 (0.49-2.31)	0.85	4/4716 (0.08%)	6/12296 (0.05%)	1.74 (0.36-7.34)	0.48
LMNA	0/4722 (0%)	0/12295 (0%)	-	-	26/4722 (0.55%)	76/12295 (0.62%)	0.89 (0.55-1.41)	0.66	23/4722 (0.49%)	34/12295 (0.28%)	1.76 (0.99-3.09)	0.038
MYBPC3	378/4604 (8.21%)	12/12295 (0.1%)	91.74 (51.72-179.29)	1.86E-198	485/4604 (10.53%)	283/12295 (2.3%)	5.00 (4.29-5.83)	1.12E-100	49/4604 (1.06%)	151/12295 (1.23%)	0.87 (0.61-1.2)	0.42
MYH7	4/4737 (0.08%)	9/12297 (0.07%)	1.15 (0.26-4.14)	0.76	517/4737 (10.91%)	152/12297 (1.24%)	9.79 (8.12-11.84)	1.28E-162	42/4737 (0.89%)	96/12297 (0.78%)	1.14 (0.77-1.65)	0.50
MYL2	4/4737 (0.08%)	5/12297 (0.04%)	2.08 (0.41-9.66)	0.28	28/4737 (0.59%)	14/12297 (0.11%)	5.22 (2.65-10.73)	1.78E-07	5/4737 (0.11%)	2/12297 (0.02%)	6.49 (1.06-68.33)	0.02
MYL3	1/4737 (0.02%)	0/12297 (0%)	-	0.28	28/4737 (0.59%)	19/12297 (0.15%)	3.84 (2.07-7.29)	5.24E-06	3/4737 (0.06%)	7/12297 (0.06%)	1.11 (0.19-4.88)	1.00

PKP2	13/4735 (0.27%)	7/12297 (0.06%)	4.83 (1.79-14.32)	0.00057	43/4735 (0.91%)	80/12297 (0.65%)	1.40 (0.94-2.06)	0.086	24/4735 (0.51%)	47/12297 (0.38%)	1.33 (0.78-2.22)	0.29
PLN	0/4735 (0%)	0/12297 (0%)	-	1	4/4735 (0.08%)	1/12297 (0.01%)	10.39 (1.03-510.59)	0.023	0/4735 (0%)	0/12297 (0%)	-	-
PRKAG2	0/4735 (0%)	3/12297 (0.02%)	0 (0-6.29)	0.57	24/4735 (0.51%)	41/12297 (0.33%)	1.52 (0.88-2.58)	0.13	16/4735 (0.34%)	34/12297 (0.28%)	1.22 (0.63-2.28)	0.53
RBM20	1/3704 (0.03%)	1/5831 (0.02%)	1.57 (0.02-123.49)	1	51/3704 (1.38%)	47/5831 (0.81%)	1.72 (1.13-2.62)	0.0089	28/3704 (0.76%)	29/5831 (0.5%)	1.52 (0.87-2.66)	0.13
SCN5A	2/4709 (0.04%)	2/12297 (0.02%)	2.61 (0.19-36.05)	0.31	83/4709 (1.76%)	222/12297 (1.81%)	0.98 (0.75-1.26)	0.90	58/4709 (1.23%)	133/12297 (1.08%)	1.14 (0.82-1.57)	0.42
TAZ	1/2878 (0.03%)	0/12281 (0%)	-	0.19	1/2878 (0.03%)	14/12281 (0.11%)	0.30 (0.01-2.00)	0.33	5/2878 (0.17%)	16/12281 (0.13%)	1.33 (0.38-3.82)	0.58
TMEM43	1/4737 (0.02%)	6/12297 (0.05%)	0.43 (0.01-3.57)	0.68	17/4737 (0.36%)	39/12297 (0.32%)	1.13 (0.60-2.05)	0.66	11/4737 (0.23%)	15/12297 (0.12%)	1.91 (0.79-4.44)	0.12
TNNC1	0/3698 (0%)	0/12297 (0%)	-	-	7/3698 (0.19%)	8/12297 (0.07%)	2.91 (0.90-9.20)	0.058	6/3698 (0.16%)	6/12297 (0.05%)	3.33 (0.89-12.46)	0.038
TNNI3	5/4673 (0.11%)	3/12180 (0.02%)	4.35 (0.85-27.98)	0.042	80/4673 (1.71%)	15/12180 (0.12%)	14.12 (8.06-26.43)	1.62E-30	4/4673 (0.09%)	19/12180 (0.16%)	0.55 (0.14-1.65)	0.35
TNNT2	8/4736 (0.17%)	5/12297 (0.04%)	4.16 (1.20-16.17)	0.011	50/4736 (1.06%)	28/12297 (0.23%)	4.67 (2.88-7.72)	3.27E-11	14/4736 (0.3%)	17/12297 (0.14%)	2.14 (0.98-4.62)	0.043
TPM1	0/4733 (0%)	0/12297 (0%)	-	-	41/4733 (0.87%)	14/12297 (0.11%)	7.67 (4.09-15.24)	7.54E-13	3/4733 (0.06%)	14/12297 (0.11%)	0.56 (0.10-2.00)	0.43
TTN	11/4730 (0.23%)	39/12297 (0.32%)	0.73 (0.34-1.46)	0.43	1136/4730 (24.02%)	2280/12297 (18.54%)	1.39 (1.28-1.51)	3.49E-15	438/4730 (9.26%)	863/12297 (7.02%)	1.35 (1.20-1.53)	1.31E-06
TTR	0/3721 (0%)	0/12297 (0%)	-	-	3/3721 (0.08%)	7/12297 (0.06%)	1.42 (0.24-6.21)	0.71	2/3721 (0.05%)	2/12297 (0.02%)	3.31 (0.24-45.62)	0.23
VCL	2/3721 (0.05%)	8/12297 (0.07%)	0.83 (0.09-4.14)	1	33/3721 (0.89%)	93/12297 (0.76%)	1.17 (0.76-1.77)	0.46	17/3721 (0.46%)	25/12297 (0.2%)	2.25 (1.14-4.35)	0.016

Table 3.6: Summary of rare variant burden testing results Analyses stratified for loss-of-function, non-truncating and synonymous variants between HCM cases (OMGL- HCMR_{EU}) and T2DM controls using a filtering allele frequency of 0.0001.

CSRP3

CSRP3 is a gene expressed in cardiac and skeletal tissue that encodes Muscle LIM protein (MLP), a protein that interacts with Z-disc proteins in sarcomeres.[189] The acronym, LIM, was assigned following the identification of a previously unreported cysteine-rich sequence common to a group of homeodomain transcription factors. This included the *Caenorhabditis elegans* cell-lineage protein *lin-11*, rat insulin gene-enhancer-binding protein *Isl1* and *Caenorhabditis elegans MEC-3*.[190–192] Consequently, the first letters of *lin-11*, *Isl1* and *MEC-3* were used to identify this as the LIM domain.[193] MLP is a small protein (194 amino acids in length), expressed in striated muscle, that appears to promote myogenic differentiation and contributes towards the structural integrity of myocytes.[189]

The finding that non-truncating *CSRP3* variants are enriched in HCM [OR: 3.10 [95% CI: 1.66 – 5.83]; p-value = 1.84×10^{-4}] supports an emerging body of evidence, including burden analyses[51], family-based studies[60] and functional studies in mice[194], that suggests *CSRP3* is a causal gene contributing towards a non-syndromic, non-sarcomeric form of hereditary HCM. When *CSRP3* was formally reviewed by ClinGen, it was deemed to have moderate evidence supporting an association with HCM.[59]

Two previous burden analyses have been performed, the first demonstrated no evidence of enrichment for non-truncating variants in *CSRP3* (HCM: (n=8/2,167) 0.37% vs. ExAC: (n=175/60,647) 0.29%; OR: 1.28 [95% CI: 0.63 – 2.60]; p-value=0.41).[8] The second burden analysis performed by Walsh et al. (2017), demonstrated enrichment in HCM cases for rare (MAF <0.0001) protein altering *CSRP3* variants by amalgamating seven previous studies together, before comparing the relative allele frequencies between HCM cases and ExAC [cases: 44/4866 (0.90%) vs. controls: 197/60706 (0.32%); p-value < 1×10^{-4}].[51] Without pursuing formal meta-analysis, and with limited details regarding the bioinformatic and variant selection process undertaken, interpretation of the magnitude of any effect size is challenging.

Burden analyses also suggest truncating variants in *CSRP3* may result in HCM, but these studies have yet to be adequately powered to definitely prove an association.[8, 51] There are case reports of homozygous truncating variants in *CSRP3* that appear to support this hypothesis.[195, 196]

Whilst burden analyses provide support for gene-level causality, not all non-truncating variants in *CSRP3* will be causal for HCM given the background rate of non-truncating *CSRP3* variants observed in controls.[8, 51, 55] Indeed, at present, only two missense variants appear to have data supporting causality, specifically *CSRP3* p.Cys58Gly and *CSRP3* p.Leu44Pro. *CSRP3* p.Cys58Gly is pathogenic based on: evidence of co-segregation (locus contained *CSRP3* yielded a maximum Logarithm of Odds (LOD) score of 5.9); supportive *in vitro* evidence of a damaging effect; absence from controls in gnomAD, and *in silico* evidence of a deleterious effect.[51, 60, 194] *CSRP3* p.Leu44Pro has evidence of enrichment in cases over controls based on burden testing (cases: 11/4866 vs. controls: 0/60706; p-value $< 1\times10^{-4}$) and supportive functional analysis performed *in vitro*.[51, 60, 194]

A murine knock-in model of *CSRP3* p.Cys58Gly demonstrates upregulation of CSRP3 mRNA and depletion of MLP, and corroborates observations from a human carrier of p.Cys58Gly.[60, 194] MLP reduction has been shown to generate a HCM phenotype in human pluripotent stem cell derived cardiomyocytes.[197] MLP reduction is thought to be attributable to chronic over-utilisation of the ubiquitin-proteasomal system, which consequentially activates bag3-mediated protein quality control processes and recruitment of numerous Hsps.[194, 198] *BAG3* is a critical regulator of cardiac proteostasis, but is traditionally described in the context of loss-of-function mutations that induce DCM or skeletal myopathy.[199–205] BAG3 contains numerous protein-protein binding domains that facilitate a panoply of functional roles, including: chaperone-assisted selective autophagy, HSP70 regulation, filamin quality control, autophagosome formation, client sequestration and ubiquitylation, client transport to aggresomes, YAP-TAZ activation and mTORC1 regulation.[206–208] Whereas downregulation of BAG3 leads to DCM, findings from the *CSRP3* mouse model suggest upregulation of *BAG3* may induce

hypertrophy signalling pathways.[194] This observation lends further support to a philosophy that HCM and DCM demonstrate diametrically opposing molecular mechanisms and represent extreme phenotypes at either end of a spectrum of myocardial function.[209] This insight may impact the long-term feasibility of gene replacement as a therapeutic strategy for individuals diagnosed with haploinsufficient *BAG3*.[208, 210]

Functional evaluation of *CSRP3* also provided insight into the role of PKCα, a protein encoded by *PRKCA* that regulates cardiac contractility in HCM.[211, 212] MLP is known to act as an endogenous inhibitor of PKCα, and depletion of MLP, as observed in the *CSRP3* p.Cys58Gly murine model leads to unopposed, chronic activation of PKCα expression.[213] However, the implications of chronic PKCα overexpression remain controversial; whilst animal studies suggest PKCα inhibition is beneficial, it contradicts findings from human genetic studies.[214]

FLNC

FLNC encodes filamin C, a large actin-crosslinking protein expressed in cardiac and skeletal muscle tissue that interacts with both the Z-disc and sarcomlemma.[215] Prior burden analysis evaluating non-truncating variants in *FLNC* was underpowered to detect relatively small effects [OR: 1.49 [95% CI: 0.75 – 2.72]; p-value = 0.203].[121] Analysis performed using the OMGL-HCMR$_{\text{EU}}$ data indicates enrichment for non-truncating *FLNC* variants in HCM cases [OR: 1.39 [95% CI:1.15 – 1.68], p-value=4.61×10^{-4}], and supports previous findings that a small proportion of non-truncating *FLNC* variants are likely to be causal of HCM.[62] For instance, a series of 448 HCM individuals from Northern Spain suggests likely pathogenic missense variants in *FLNC* account for up to ∼2% (1.34% [95% CI: 0.62 – 2.89%] (n=6/448)) of all HCM cases.[216] There was inadequate power to evaluate truncating variants in *FLNC*, although most evidence suggests that premature truncations in *FLNC* result in non-HCM forms of cardiomyopathy.[217–220]

TTN

***TTN* in dilated cardiomyopathy** *TTN*, the largest human protein, has demonstrated numerous contributions towards both the structure and function of striated muscle, with truncating variants in *TTN* (*TTN*tv) established as a prominent genetic risk factor for the development of DCM.[8, 148] Truncating variants in *TTN* occur reasonably frequently in the general population ($\sim$ 1 in 230 individuals), and are associated with a subnormal ejection fraction (OR:9.3 [95% CI: 3.9 - 22.2]; p-value = 5.7×10^{-5}) and an increased risk of adverse outcomes, specifically incident DCM, heart failure, or all-cause mortality (HR: 2.5; [95% CI 1.4 - 4.3]; p-value = 1.1×10^{-3}).[125, 221] Experimental evidence indicates that *TTN*tv lead to haploinsufficiency.[125] Whereas the role of *TTN*tv in DCM is relatively well documented, *TTN* – and more specifically non-truncating and synonymous variants in *TTN* – is less well understood in the context of HCM. *TTN* missense variants do not appear to be enriched in DCM cases, and similarly *TTN*tv are not enriched in HCM cases, as validated in the OMGL- HCMR_{EU} data.[147, 222]

***TTN* in hypertrophic cardiomyopathy** Prior burden testing analysis, performed in the BRRD cohort, indicated that non-truncating *TTN* variants were not enriched in HCM cases (OR:0.91 [95% CI: 0.69-1.20]; p-value=0.55).[121] However, examination of this prior approach, performed in the BRRD cohort, indicates that the analysis was not suitably powered to detect relatively small effect sizes and was not specific to *TTN* exons that are deemed to be clinically informative (i.e. PSI >0.9).[121] Given this context, and the finding that HCM cases from the OMGL- HCMR_{EU} cohort possess a significant excess of non-truncating (OR: 1.39 [95% CI:1.28-1.51]; p-value=3.49×10^{-15}) and synonymous variants (OR:1.35 [95% CI: 1.20 – 1.53]; p-value=1.31×10^{-6}) in *TTN*, this supports the hypothesis that non-truncating and synonymous variants contribute towards the genetic architecture of HCM.

Alternatively, it could be hypothesised that enrichment for non-truncating and synonymous *TTN* variants in the OMGL- HCMR_{EU} cohort are reflective of

systematic or *TTN*-specific features. Systematic issues that may have contributed towards these observations would include inherent differences between cases and controls, such as ancestral differences. Further analysis was performed to assess whether population stratification might explain the enrichment of non-truncating and synonymous *TTN* variants in the OMGL- $HCMR_{EU}$ cohort. The OMGL cohort were excluded, as their underlying ancestry remains relatively unknown, and analysis was limited to only those individuals deemed, on average to be genetically similar to European individuals from the 1000 Genomes phase 3 project and not closely related to one another, from either the HCMR or T2DM cohorts. However, an excess of non-truncating ($HCMR_{EU}$:24.1% (476/1974) vs. $T2DM_{EU}$:18.5%; OR: 1.40 [95% CI:1.24-1.56]; p-value=1.27×10^{-8}) and synonymous variants ($HCMR_{EU}$:9.32% (184/1974) vs. $T2DM_{EU}$:7.02%; OR:1.36 [95% CI: 1.15 – 1.61]; p-value=4.01×10^{-4}) in *TTN* persisted in HCM cases. Other systematic issues that could be speculated as a potential cause include the use of multiple different sequencing platforms and chemistries, which could not be adjusted for. Introduction of a technical artefact through this mechanism would have implications for other genes in the analysis, and given that most findings emerging from these data are confirmatory of prior findings, and no excesses of synonymous variants were detected across other genes, this scenario is unlikely.

It is also possible that the observed excess of non-truncating and synonymous variants in *TTN* are attributable to challenges in processing *TTN* itself. *TTN* is extremely large, (encodes 27-33k amino acids), and has known difficulty with variant calling given its size, and the presence of sequence homology.[223] It is possible that a subset of synonymous variants in *TTN* do not confer neutral effects and are instead cryptic splice sites; and bioinformatic tools, such as TraP and SpliceAI, could be deployed to further evaluate this.[224, 225]

However, ultimately to determine whether non-truncating and synonymous variants in *TTN* contribute towards the genetic architecture of HCM, replication will be required from a large, independent case-control analysis.

Non-truncating *TTN* variants in hypertrophic cardiomyopathy HCM and DCM can both result from disease causing variants in sarcomere proteins, but variants associated with HCM confer opposing calcium sensitivities to those associated with DCM, suggesting the molecular mechanisms underpinning these two conditions are diametrically opposing one another. As a corollary, whereas *TTN*tv lead to *TTN* loss-of-function, it could be hypothesised that non-truncating *TTN* variants contribute towards *TTN* gain-of-function, with a poison peptide restricting wild-type sarcomere functionality. Whilst rare *TTN* missense variants are encountered relatively frequently, there are very few examples where pathogenic *TTN* missense variants have been identified. ClinVar reports *TTN* p.Arg740Leu as a pathogenic variant, but with supporting evidence derived from a 1999 study that failed to assess the frequency of this specific allele in 260 unrelated healthy controls, it is unlikely that this remains a pathogenic variant.[226] More recently, a *TTN* missense variant (p.Lys22368Asn) has been classified as pathogenic by the authors of a case report that described *TTN* p.Lys22368Asn in the context of a possible oligogenic cause of HCM.[72] Furthermore, a common missense variant in *TTN*, rs2042995 (effect allele = T, minor allele frequency ∼77% in gnomAD Europeans), has been reported to contribute towards cardiac magnetic resonance derived traits underpinning myocardial structure and function, specifically: left ventricular end diastolic volume (OR 1.09 [95% CI:1.07-1.12]; p-value=2.30×10^{-11}), left ventricular end systolic volume (OR 1.13 [95% CI:1.10-1.15]; p-value=8.40×10^{-20}) and left ventricular ejection fraction (OR 0.91 [95% CI:0.89-0.94]; p-value=2.50×10^{-12}).[227] Collectively, it could be hypothesised that non-truncating variants in *TTN* have a role in modifying HCM expressivity.

Synonymous *TTN* variants in hypertrophic cardiomyopathy As a group, synonymous variants are presumed to have a neutral effect on disease risk. Central to this assumption is the belief that whilst synonymous variants result in transcription of an alternative mRNA codon, they do not alter the amino acid that is ultimately transcribed, as a result of codon degeneracy. In humans, synonymous variants

do not appear to be under a selection pressure, as evidenced by the observation that synonymous variants are almost randomly distributed across the genome, which contrasts from observations in Drosophila melanogaster.[228, 229] There is evidence to suggest synonymous variants can directly cause human disease, generally through disrupted splicing.[230–232] Other mechanisms have also been postulated, including how non-random usage of synonymous codons can disrupt exonic transcription factor binding, influence mRNA stability, protein expression, protein conformation and protein function.[233–236] Many of these mechanisms are a result of an elongated translation phase. Sharp et al. (1986) attempted to quantify this metric through development of the "relative synonymous codon usage" metric in yeast, and it has been subsequently applied to many synonymous variants associated with human diseases.[233, 237]

The observation that synonymous *TTN* variants are enriched in HCM cases (OR:1.35 [95% CI: 1.20 – 1.53]; p-value=1.31×10^{-6}) stimulates several further hypotheses. Whether or not the small effect size is attributable to a large number of synonymous variants, each having a small effect size, or a small number of synonymous variants, each with a large effect but present in the context of a large number of synonymous variants of neutral effect, that consequently dilutes the signal, is yet to be established. Synonymous variant codon bias in *TTN* has not been formally evaluated, but may deserve further evaluation.

3.4 Discussion and limitations

Combining gene panel sequencing data from the OMGL and HCMR cohorts allowed the monogenic architecture of HCM to be further explored. Conducted in a large case-series, exposed to rigorous quality control procedures, the findings here are largely confirmatory of prior knowledge, and therefore provide reassurance in the methodology that has been implemented in the generation of these data.

The OMGL-HCMR cohort supports dogma that a minority of individuals yield disease-causing variants, with 39.3% [95% CI: 38.0-40.6%] (n=2,117/5,393) harbouring a pathogenic, likely pathogenic or variant of uncertain significance

across 35 cardiomyopathy associated genes, with truncating variants in *MYBPC3* and missense variants in *MYH7* the leading causes. *MYBPC3* c.1224-52G>A, an intronic variant that disrupts splicing, was found to be a frequent cause of HCM (accounts for 1.04% [95% CI:0.80 – 1.35] of HCM), and suggests clinical diagnostic gene panel sequencing should extend beyond the traditional exonic boundaries to safeguard against missing additional pathogenic variants. In the first instance, diagnostic tests could aim to incorporate intronic regions across genes known to act through loss-of-function mechanisms to try and capture variants that introduce cryptic splice sites. In practice, this recommendation may be difficult to implement as many clinical diagnostic labs undertake exome sequencing, for which capture regions are predefined and potentially difficult to change. Consequently, the evidence provided here could be used to support the use of genome sequencing as a clinical diagnostic test in circumstances where a rare disease-causing variant has not been identified via exome sequencing. An advantage to this approach is that up to 40% of probands will receive a genetic diagnosis following exome sequencing, and genome sequencing can be then be used to capture non-coding regions (particularly those that have the potential to impact gene product formation, such as introns and 5' or 3' UTR regions), missed by exome sequencing, in the remaining 60% of probands.

Before the introduction of contemporary variant classification methods the prevalence of multiple disease-causing variants in HCM was believed to be ∼8%, but data generated from the OMGL-HCMR cohort replicate more recent literature findings that suggest this is a rare phenomena, isolated to ∼1:1000 HCM cases.[176] However, without clinical correlates in the OMGL-HCMR cohort, strongly held assumptions that individuals with multiple disease-causing variants experience a more severe phenotype and worse prognosis could not be formally explored.

Burden testing indicated that the core sarcomere genes were enriched for rare missense variants in HCM cases, alongside the enrichment of predicted loss-of-function variants in *MYBPC3*. Furthermore, enrichment for missense variants was noted in *CSRP3* and *FLNC*, providing supportive evidence regarding the role of these genes in HCM. Missense variants in *TTN* were also found to be enriched

in HCM cases, but this association remains largely speculative, with no obvious biological mechanism and potentially confounded by the observation that HCM cases demonstrated an excess of *TTN* synonymous variants.

4
Oligogenicity

Contents

4.1 Background

Traditionally, the genetic architecture of disease was explained using a power law curve, considering both the population prevalence of a condition and the effect sizes conferred by variants underpinning those diseases.[109, 238] Supporting this viewpoint was the observation that monogenic diseases were rare and disease-causing variants associated with these diseases were highly penetrant. Conversely,

polygenic diseases, such as type two diabetes or hypertension, were relatively common and susceptibility towards developing these conditions was underpinned by many variants, each of relatively small effect (the common disease common variant hypothesis).[239] It was hypothesised that oligogenicity might partially explain the genetic architecture of diseases that neither demonstrated monogenic nor polygenic inheritance patterns.[102, 104] Oligogenicity is defined by the presence of multiple variants across two or more candidate genes that independently would be insufficient to cause disease, but in aggregate are both sufficient and necessary to result in disease.[102] As such, oligogenicity is assumed to result from gene-gene interactions, otherwise referred to as epistasis.[102] Despite relatively few established oligogenicity examples existing, the DIDA database has been developed to document suspected oligogenicity patterns (http://dida.ibsquare.be/).[240]

The genetic architecture of sarcomere-negative HCM has previously been considered and oligogenicity postulated as a possible mechanism, particularly for individuals presenting with a milder HCM phenotype without a family history of disease.[2, 241, 242] In support of this hypothesis was the finding of *MYBPC3* c.3628-41_3628-17del (*MYBPC3*Δ25), a variant that induces a 25 base pair deletion within intron 32. Despite being present in 4% of South Asian individuals, this variant results in a partial disruption of *MYBPC3* splicing (rather than complete splicing disruption that would have been expected with monogenic HCM).[2, 107] In a proportion of *MYBPC3*Δ25 carriers it has been hypothesised that additional low-penetrant variants act, presumably at a pathway-level, and confer heightened susceptibility towards HCM.[107, 243–247] However, it remains to be proven whether sarcomere negative individuals are enriched for multiple low-penetrant variants across the core sarcomere genes. Studies to date have instead focused on presenting case-only analysis, without considering the individual-level distribution of variants in an unaffected population to gauge whether what is observed in cases is to be expected.[72]

The key objectives for this Chapter are to:

1. Evaluate the distribution of rare variants across core sarcomere genes in cases and controls.

2. Perform a conditional analysis to evaluate for additional low-penetrant variants acting in combination with MYBPC3Δ25

4.2 Distribution of rare variants across core sarcomere genes

4.2.1 Demographic details

Analysis was performed using the BRRD case-control cohort. The BRRD includes 214 sarcomere negative HCM cases and 5,802 controls, distantly related to one another, and using PCA, deemed to be genetically representative of European individuals enrolled through the 1000 Genomes phase 3 project (Chapter 2). Demographic details are summarised in Tables 3.1 and 3.2. The BRRD is potentially well suited to study oligogenicity. With cases and controls sequenced collectively on the same platform, technical artefacts are minimised. Furthermore, the BRRD provides access to individual-level sequence data, rather than summary-level data, for all participants, which is essential for the evaluation of multiple variants.

4.2.2 Evidence supporting a multiple variant hypothesis

Missense and synonymous variants, with a minor allele frequency in the gnomAD non-Finnish European population < 0.01, were extracted from eight core sarcomere genes (*ACTC1, MYBPC3, MYH7, MYL2, MYL3, TNNI3, TNNT2, TPM1*) present in BRRD. These eight core sarcomere genes were selected for exploratory analysis as mutations detected in these genes are well established causes of HCM, and this helps facilitate interpretation of any novel results. The number of variants, stratified by VEP-assigned consequence from across a range of minor allele frequency thresholds (<0.01, <0.001, <0.0001 alongside 0.01-0.001 and 0.001-0.0001), that were harboured amongst 8 core sarcomere genes per individual were counted. The specific VEP-assigned consequences evaluated included a composite group of

protein altering variants (composed of missense, stop gained, frameshift, splice donor/acceptor/region or inframe insertion/deletion variants), in addition to a group of missense variants, and a group of synonymous variants. Fisher's exact test was employed to establish whether there were differences, between cases and controls, regarding the number of variants carried per individual, due to the presence of several counts with low or absent counts. First, an aggregate burden test was performed to assess whether there was an excess aggregate burden of variants in HCM cases across the 8 core sarcomere genes. Second, differences between the number of individuals harbouring multiple (i.e. greater than or equal to two) variants across 8 core sarcomere genes was assessed.

Aggregate burden

This exploratory analysis was designed to assess the proportion of cases and controls that harboured at least 1 rare variant across 8 sarcomere genes. Synonymous variants appeared balanced between groups (Table 4.1). There was evidence that rare (MAF <0.0001) protein-altering variants are enriched in HCM cases, but this may reflect the presence of individuals with a disease-causing variant, yet to be formally defined as likely pathogenic/pathogenic (i.e. a VUS).

Variant type	MAF threshold	Counts				Proportions		RR	≥1 variants		
		Cases <1	Cases ≥1	Controls <1	Controls ≥1	Cases	Controls		OR	P value	Power
Synonymous	<0.01	194	20	5189	613	0.09	0.11	0.82	0.87 [0.51-1.40]	0.65	0.13
	0.01-0.001	198	16	5439	363	0.07	0.06	1.17	1.21 [0.67-2.04]	0.472	0.11
	<0.001	210	4	5520	282	0.02	0.05	0.40	0.37 [0.10-0.98]	0.047	0.56
	0.001-0.0001	213	1	5684	118	0.005	0.02	0.25	0.23 [0.005-1.30]	0.131	0.31
	<0.0001	211	3	5634	168	0.01	0.03	0.33	0.48 [0.097-1.44]	0.29	0.16
Protein-altering	<0.01	183	31	5080	722	0.14	0.12	1.17	1.19 [0.78-1.77]	0.399	0.15
	0.01-0.001	204	10	5489	313	0.05	0.05	1.00	0.86 [0.40-1.64]	0.758	0.04
	<0.001	190	24	5382	420	0.11	0.07	1.57	1.62 [1.00-2.52]	0.044	0.56
	0.001-0.0001	207	7	5622	180	0.03	0.03	1.00	1.06 [0.413-2.26]	0.84	0.06
	<0.0001	194	20	5549	253	0.09	0.04	2.25	2.26 [1.33-3.66]	0.002	0.85
Missense	<0.01	189	25	5262	540	0.12	0.09	1.33	1.29 [0.81-1.98]	0.233	0.26
	0.01-0.001	209	5	5590	212	0.02	0.04	0.50	0.63 [0.20-1.52]	0.452	0.1
	<0.001	193	21	5468	334	0.1	0.06	1.67	1.78 [1.06-2.85]	0.018	0.6
	0.001-0.0001	207	7	5654	148	0.03	0.03	1.00	1.29 [0.50-2.78]	0.505	0.14
	<0.0001	198	16	5608	194	0.07	0.03	2.33	2.34 [1.28-3.99]	0.004	0.77

Table 4.1: Aggregate burden of rare variants across sarcomere genes Exploratory analysis used to evaluate the proportion of individuals who harbour at least 1 rare variant across 8 sarcomere genes. Synonymous group representative of a negative control. The relative risk (RR) denotes the relative proportional difference between cases and controls. P-values highlighted in red are significant following Bonferroni correction ($\alpha = 0.05/15$). Power highlighted in yellow is greater than 80%.

Multiple variants per individual

The relatively low proportion of individuals in whom at least one rare variant was detected restricted more nuanced analyses pertaining to oligogenicity. As a general exploratory analysis, the relative proportion of cases and controls carrying two or more rare variants across 8 core sarcomere genes was evaluated. This is not strictly assessing oligogenicity, as it encompasses both oligogenic and double heterozygous mechanisms. Nevertheless, following Bonferroni correction to limit false-positive associations, there was no evidence of enrichment for protein-altering or missense variants in sarcomere negative cases (Table 4.2). However, there is evidence of nominal significance for protein-altering variants (MAF <0.01). It is therefore possible that the multiple variant analysis was not adequately powered to formally reject the null hypothesis (Table 4.2).

Variant type	MAF threshold	Counts				Proportions		RR	≥2 variants		
		Cases <2	Cases ≥2	Controls <2	Controls ≥2	Cases	Controls		OR	P value	Power
Synonymous	<0.01	213	1	5756	46	0.005	0.008	0.63	0.59 [0.01-3.48]	1	0
	0.01-0.001	213	1	5794	8	0.005	0.001	5.00	3.40 [0.08-25.5]	0.278	0.17
	<0.001	214	0	5794	8	0	0.001	0	0 [0-16.0]	1	0
	0.001-0.0001	214	0	5801	1	0	0.0002	0	0 [0-1040]	1	0
	<0.0001	214	0	5799	3	0	0.001	0	0 [0-65.8]	1	0
Protein-altering	<0.01	209	5	5761	41	0.023	0.007	3.29	3.36 [1.03-8.62]	0.023	0.54
	0.01-0.001	214	0	5796	6	0	0.001	0	0 [0-23.1]	1	0
	<0.001	211	3	5777	25	0.014	0.004	3.50	3.28 [0.63-10.9]	0.076	0.41
	0.001-0.0001	214	0	5801	1	0	0.0002	0	0 [0-1040]	1	0
	<0.0001	214	0	5789	13	0	0.002	0	0 [0-8.94]	1	0
Missense	<0.01	211	3	5780	22	0.014	0.004	3.50	3.73 [0.71-12.6]	0.057	0.48
	0.01-0.001	214	0	5800	2	0	0.0003	0	0 [0-145]	1	0
	<0.001	212	2	5788	14	0.009	0.002	4.50	3.90 [0.43-17.1]	0.109	0.27
	0.001-0.0001	214	0	5801	1	0	0.0002	0	0 [0-1040]	1	0
	<0.0001	214	0	5795	7	0	0.001	0	0 [0-18.9]	1	0

Table 4.2: Presence of multiple rare variants across 8 sarcomere genes Exploratory analysis used to evaluate the proportion of individuals who harbour at least 2 rare variant across 8 sarcomere genes. Synonymous group representative of a negative control. The relative risk (RR) denotes the relative proportional difference between cases and controls. P-values highlighted in red are significant following Bonferroni correction ($\alpha = 0.05/15$). Power highlighted in yellow is greater than 80%.

4.2.3 Power calculation

Power calculations were performed in R using the *statmod* package (https://cran.r-project.org/package=statmod). Simulated case-control data was generated to assess the relationship between power and sarcomere negative cohort size, assuming a fixed control cohort of 5,000 individuals. The proportion of controls harbouring two or more variants was derived from empirical BRRD data, specifically: 2×10^{-4}, 0.002, 0.004 and 0.008 (Table 4.2). Experimental power is also contingent on the relative difference in carrier proportions between cases and controls, and represented by the relative risk (RR). For example, if the proportion of controls with at least 2 variants is 0.008, a RR of 1 would indicate the proportion of cases would also be 0.008, whereas a RR of 2 would suggest the case proportion would be 0.016.

Under the above assumptions, a power calculation was performed (Figure 4.1). To achieve 80% power, assuming a RR of 2 and control proportion of 0.008, a case-control study of ∼2,000 sarcomere negative cases and 5,000 controls would be required. Empirical BRRD data indicates a control proportion of 0.008 is only achievable for the protein-altering variant group with a MAF threshold of <0.01. If the RR were ∼3.5, as observed in the BRRD data, it is possible that 80% power could be achieved with ∼1000-1500 cases assuming a control proportion of between 0.002 and 0.004.

4.3 Low-penetrant variants and *MYBPC3*Δ25

4.3.1 The role of *MYBPC3*Δ25 in oligogenicity

Quantifying the oligogenic burden of low-frequency variants in HCM was not possible due to limited discovery power in the BRRD cohort. However, prior analyses suggest *MYBPC3*Δ25 is a contributor towards an oligogenic model of HCM, and this specific hypothesis was formally evaluated. First, evidence supporting the role of *MYBPC3*Δ25 in HCM will be reviewed, before specific hypotheses relating to *MYBPC3*Δ25 are formally tested.

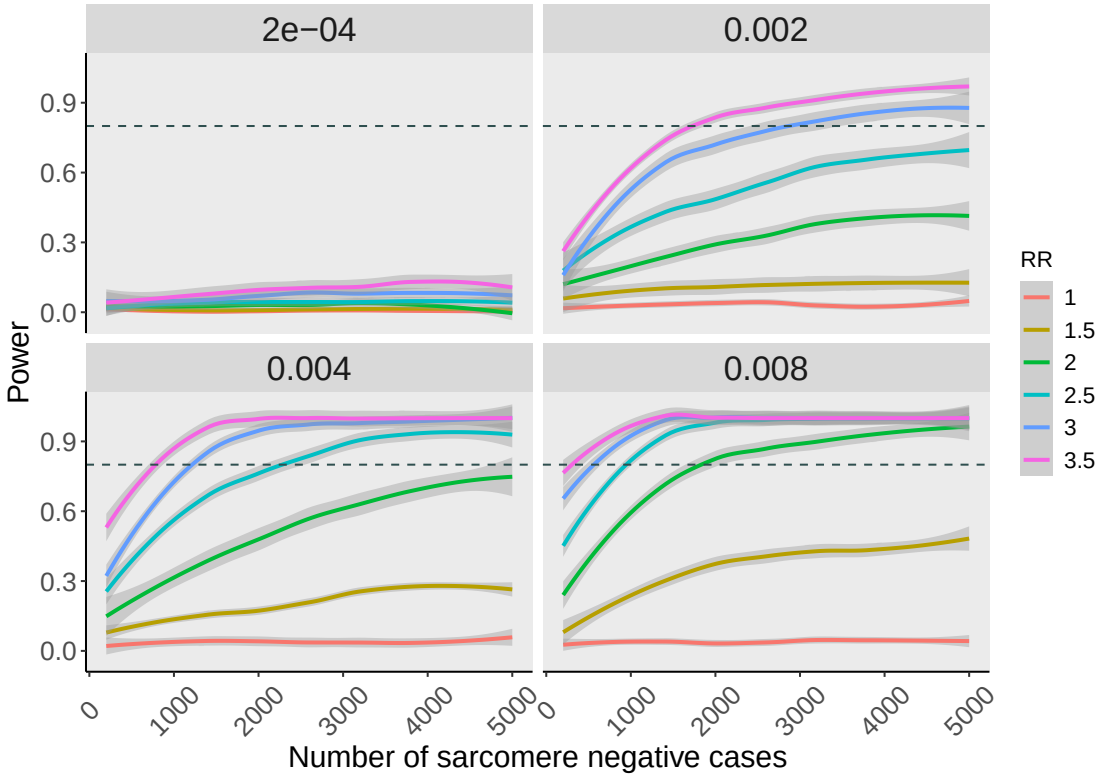


Figure 4.1: Power calculation assessing multiple variant enrichment in HCM cases Figure constructed using simulated case-control data (200-5000 cases vs. 5000 controls) accommodating the relative risk (RR) between cases and controls (between 1 and 3.5) and the proportion of controls found to harbour two or more rare variants.

4.3.2 *MYBPC3*Δ25

MYBPC3 c.3628-41_3628-17del, causes a 25 base pair deletion within intron 32 (*MYBPC3*Δ25) and results in partial splicing. *MYBPC3*Δ25 is considered an ancestry specific variant, having been detected almost exclusively in individuals of South Asian ancestry (in gnomAD v2.1.1 98.1% [95% CI: 97.0–98.9%] of *MYBPC3*Δ25 variant carriers are of South Asian ancestry). In gnomAD (v2.1.1), 6.16% of individuals ascribed South Asian ancestry were heterozygous for the *MYBPC3*Δ25 variant (943/15,296, [95% CI 5.75-6.50%]) and 0.12% were homozygous (n=19). This is consistent with previous studies which have reported carrier frequencies ranging from 2 to 8% in individuals of South Asian ancestry.[107, 248] Estimates suggest there are approximately 100 million heterozygous carriers of *MYBPC3*Δ25 globally, with *MYBPC3*Δ25 reported to confer an almost 7-fold increased risk of

cardiomyopathy.[107] However, the original analysis presented by Dhandapany et al. (2009), used two heterogeneous composite case groups consisting of individuals diagnosed with HCM (n=357), DCM (n=395), and restrictive cardiomyopathy (RCM) (n=15).[107] Given the diametrically opposing molecular mechanisms that underpin sarcomeric HCM and DCM, the rationale for including multiple subtypes is uncertain and makes interpretation of *MYBPC3*Δ25 in the context of any one of these cardiomyopathies more challenging.[249–251] Given that truncating variants in *MYBPC3* have only been associated with HCM, and not primary DCM, it is reasonable to assume that *MYBPC3*Δ25 is of most relevance to HCM, but the true effect size may be underestimated.[8]

The ACMG/AMP variant classification framework indicates that only variants with an allele frequency exceeding 5% are considered benign based on allele frequency alone, as specified by the BA1 rule. Evaluation in gnomAD (v2.1.1) indicates that *MYBPC3*Δ25 has an allele frequency of 0.0321 [95% CI: 0.0302 – 0.0341] (allele count=981, allele number=30,592) in South Asian individuals, and whilst this does not automatically provoke the BA1 rule, it is in excess of the credible allele frequency threshold (i.e. 1×10^{-4}) that has been established for disease-causing variants in HCM (Chapter 2). In opposition to the hypothesis that *MYBPC3*Δ25 is benign, numerous reports exist on ClinVar that document *MYBPC3*Δ25 as a disease-causing variant, and *in vitro* functional studies have previously suggested *MYBPC3*Δ25 induces exon skipping of exon 33 by removing the splice branch point in intron 32.[107, 252] *MYBPC3*Δ25 does not, therefore, align with a monogenic model of HCM, whereby a single variant incurs a large effect size, but instead appears more likely to represent a low-penetrant variant and may provide insight into a possible oligogenic model of HCM. However, alternative hypotheses must also be considered, including the possibility that *MYBPC3*Δ25 is in LD with a rarer, more penetrant variant, which pertains to the concept of synthetic association.[253–255]

A synthetic association hypothesis has been partially explored in South Asian carriers of *MYBPC3*Δ25 previously.[248] But, importantly, this previous study was conducted in the general population, rather than in individuals with a firm clinical

diagnosis of HCM. Specifically, Viswanathan et al. (2018) detected a rare variant, *MYBPC3* p.D389V, that exclusively occurred on the *MYBPC3*Δ25 haplotype, and carriers of *MYBPC3*Δ25bp/D389V demonstrated hyperdynamic echocardiographic features suggestive of early stage HCM.[248] Conclusions drawn from the study are limited given that the analysis was not performed in HCM cases, and *MYBPC3* p.D389V has not yet been publicly reported in a HCM case. Nevertheless, the hypothesis that synthetic association occurs between *MYBPC3*Δ25 and a rare disease-causing variant (Figure 4.2) is valid and will be assessed, prior to pursuing analyses addressing the potential co-occurrence of multiple low-penetrant variants in sarcomere-negative HCM cases.

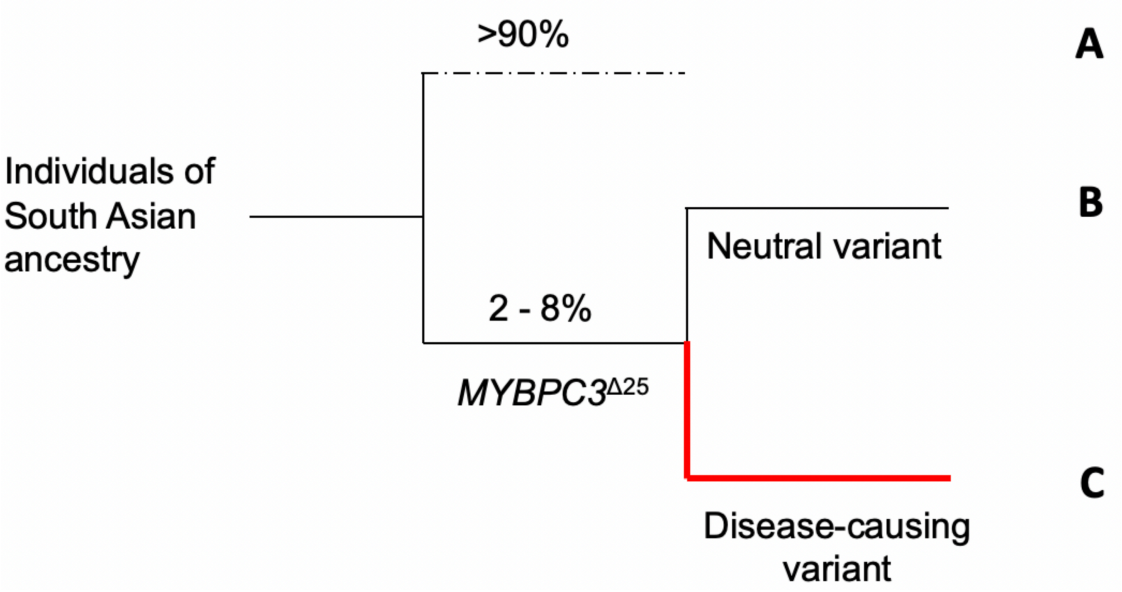


Figure 4.2: Simplified view of an ancestral tree to demonstrate a hypothesis relating to synthetic association between *MYBPC3*Δ25 and an underlying disease-causing variant in individuals of South Asian ancestry. In this schematic there are three possible outcomes (A, B and C). Outcome A will be applicable to the majority of individuals, as they do not possess *MYBPC3*Δ25. Outcomes B and C relate specifically to carriers of *MYBPC3*Δ25, with individuals diagnosed with HCM (highlighted in red) likely to carry a disease-causing variant that has arisen after the *MYBPC3*Δ25 (Outcome C). For individuals without a HCM diagnosis, but proven to carry *MYBPC3*Δ25, I would anticipate no disease-causing variants to be detected (Outcome B).

4.3.3 *MYBPC3*Δ25 in hypertrophic cardiomyopathy cases

In the HCMR cohort, 0.68% of individuals (n=18/2,636) were heterozygous for the *MYBPC3*Δ25 variant. No homozygous *MYBPC3*Δ25 carriers were detected in either cohort. This observation aligns with the expected prevalence of homozygous carriers for *MYBPC3*Δ25, as reported in gnomAD. A 2-sample test for equality of proportions with continuity correction shows no significant difference (p-value >0.05). As the HCMR cohort had undertaken genome-wide genotyping, individuals could be assigned ancestral groups based on principal components projected onto the 1000 Genomes Phase 3 data. 12.7% [95% CI 8.1-19.4%] (n=17/134) of individuals ascribed South Asian ancestry by PCA were carriers of *MYBPC3*Δ25, aligning with observations from gnomAD that *MYBPC3*Δ25 is almost exclusively detected in individuals of South Asian ancestry. In the HCMR cohort, 94% [95% CI:74.2-99.7%] (n=17/18) *MYBPC3*Δ25 variant carriers were deemed to be of South Asian ancestry by PCA. A similar proportion of the OMGL HCM cohort, 0.73% [95% CI: 0.47 – 1.1%] (n=20/2,757), were found to be heterozygous carriers for the *MYBPC3*Δ25 variant. It was not possible to identify the total number of South Asian individuals within the OMGL cohort through PCA, as genome-wide genotyping was not performed. However, if we assume the OMGL cohort is proportionately representative of the United Kingdom, in terms of ancestry, and that *MYBPC3*Δ25 is specific to South Asian ancestry, we can deduce that ∼4.8% of individuals included in the OMGL cohort are of South Asian ancestry (n=132). This is calculated based on 2011 census data provided by the Office for National Statistics (https://www.ons.gov.uk/) that suggests 3,078,374 individuals are of Indian, Bangladeshi or Pakistani ethnicity, from a total population of 63,182,178. Based on these assumptions, approximately 15.1% [95% CI: 10.0 – 22.2] (n=20/132) of individuals presumed to demonstrate South Asian ancestry are carriers for *MYBPC3*Δ25, similar to what is reported in the HCMR cohort. Using similar logic, it is assumed that across the United Kingdom there are ∼100,000 carriers of *MYBPC3*Δ25, a striking estimate given that the total number of individuals

affected by HCM in the United Kingdom is approximated to be ∼130,000 (given HCM prevalence of 1:500 and total population of 63 million).

HCM probands of South Asian ancestry from the HCMR cohort demonstrated detection rates for disease causing variants (25.4% [95% CI: 18.8 – 33.4] (n=34/134) for pathogenic/likely pathogenic variants and 15.6% [95% CI: 10.5 – 22.8] (n=21/134) for VUSs) comparable to those observed in the presumed mixed ancestry OMGL cohort, and previously published cohorts.[8, 55] If a monogenic disease model were to be assumed, the penetrance associated with *MYBPC3*Δ25 is very low, and approximated to be 0.002 [95% CI: 0.0014-0.0029] for all ancestral groups. Refining the calculation to only consider individuals of South Asian ancestry makes a small difference to the overall penetrance estimate (0.0087 [95% CI: 0.006-0.012]. The equation used to calculate this penetrance estimate has been extracted from Minikel et al (2016) (Figure 4.3).[256]

$$Penetrance = population\ prevalence \times \frac{case\ allele\ frequency}{population\ control\ allele\ frequency}$$

Figure 4.3: Penetrance equation Penetrance was calculated using the following parameters: population prevalence = 1/500; all-ancestry case allele frequency = case allele count (n=38) / case allele number (n = 10,786); South Asian case allele frequency = case allele count (n=37) / case allele number (n = 266); all-ancestry population control allele frequency = population allele count (n=1000) / observed population allele number (n=279,908); South Asian population control allele frequency = population allele count (n=981) / observed population allele number (n=30,592). The allele count represents the number of times *MYBPC3*Δ25 is observed. The allele number denotes the total number of chromosomes that have been evaluated for *MYBPC3*Δ25. Observed population allele counts and allele numbers were extracted from gnomAD v2.1.1.

4.3.4 Disease-causing variants accompanying *MYBPC3*Δ25 in HCM cases

To formally evaluate the hypothesis that HCM disease risk is attributable to a rare variant in LD with *MYBPC3*Δ25, rather than the independent effects of *MYBPC3*Δ25, a case-control study was conducted. Firstly, individuals with HCM,

derived from the OMGL and HCMR cohorts, were identified who carried both *MYBPC3*Δ25 and an additional disease-causing variant (Table 4.3).

Pathogenic or likely pathogenic sarcomere gene variants were detected in 50% [95% CI: 34.8 – 65.2%] (n=19/38) of individuals heterozygous for *MYBPC3*Δ25 across the OMGL and HCMR cohorts. The most frequently observed pathogenic or likely pathogenic variant, found in 28.9% [95% CI: 17.0 – 44.8%] (n=11/38) of individuals heterozygous for *MYBPC3*Δ25, was *MYBPC3* c.1224-52G>A. *MYBPC3* c.1224-52G>A is a recognised pathogenic variant.[74] *MYBPC3* c.1224-52G>A is frequently observed in both the OMGL cohort (32 of 2,757 probands (1.2% [95% CI: 0.8-1.6%])) and HCMR cohort (23 of 2,636 probands(0.9%, [95% CI: 0.6-1.2%])). Variants of uncertain significance accompanied *MYBPC3*Δ25 in five individuals.

4.3.5 Evaluating the relationship between *MYBPC3*Δ25 and *MYBPC3* c.1224-52G>A

Using gene panel sequence data, specific to individuals of South Asian genetic ancestry from the HCMR cohort (n=134), haplotype analysis was conducted using Haploview (version 4.2). Genetic markers located within MYBPC3 with a minor allele frequency > 0.01 within this population, minimum genotyping rate of 90%, HWE p-value cut-off of 0.001 and a maximum of one Mendel error, were considered alongside *MYBPC3* c.1224-52G>A and *MYBPC3*Δ25. There is evidence of strong LD between *MYBPC3*Δ25 and *MYBPC3* c.1224-52G>A ($D' = 0.81$ and $r^2 = 0.22$) (Figure 1). Using predicted haplotype frequencies from an LD model the allelic association between *MYBPC3*Δ25 and *MYBPC3* c.1224-52G>A corresponds to an estimated OR of 98.0 [95% CI: 9.92 – 4763.71; p-value=3.22×10^{-6}] (Table 4.4). The *MYBPC3* c.1224-52G>A variant was found to occur on the second most commonly observed *MYBPC3*Δ25 haplotype (Figure 4.5). Direct comparison of the proportion of heterozygous *MYBPC3*Δ25 variant carriers between the HCMR (n=17/134) and gnomAD (n=943/15,296) South Asian cohorts indicated a 2-fold enrichment within HCM cases, (OR: 1.98 [95% CI 1.11-3.50], p-value=0.015) (Table 4.5). When HCMR probands with the *MYBPC3*Δ25/-52 haplotype were excluded, no difference

Gene	Variant	Variant classification	Frequency in individuals heterozygous for MYBPC3Δ25bp	
			OMGL (n=20)	HCMR (n=18)
MYBPC3	c.1224-52G>A	Pathogenic	6	5
MYBPC3	c.1227-13G>A	Pathogenic	1	1
MYBPC3	c.2827C>T p.(Arg943Ter)	Pathogenic	1	-
MYH7	c.2770G>A p.(Glu924Lys)	Pathogenic	1	-
MYBPC3	c.821+2T>C	Pathogenic	-	1
MYH7	c.1988G>A p.(Arg663His)	Pathogenic	-	1
MYH7	c.2221G>A p.(Gly741Arg)	Pathogenic	-	1
MYBPC3	c.2308G>A p.(Asp770Asn)	Likely pathogenic	1	-
MYBPC3	c.2030C>T p.(Pro677Leu)	VUS	1	-
MYH7	c.3931C>G p.(Gln1311Glu)	VUS	1	-
MYH7	c.436A>G p.(Lys146Glu)	VUS	1	-
MYH7	c.5065C>T p.(Arg1689Cys)	VUS	-	1
MYH7	c.170G>A p.(Gly57Asp)	VUS	-	1

Table 4.3: Pathogenic, likely pathogenic and variants of uncertain significance accompanying *MYBPC3*Δ25 in individuals from both the OMGL and HCMR cohorts. NCBI transcript IDs: *MYBPC3*: NM_000256.3, *MYH7* NM_000257.2. ACMG criteria applied in classification

was observed (OR: 0.96 [95% CI 0.40-1.95], p-value=1.0). However, this approach does not directly address whether *MYBPC3*Δ25 or *MYBPC3* c.1224-52G>A is responsible for HCM risk. Therefore to model the independent effects of the *MYBPC3*Δ25 and *MYBPC3* c.1224-52G>A variants on HCM risk, whilst allowing for LD between the two variants, South Asian cases and controls, defined using PCA,

were directly evaluated: cases were derived from the HCMR cohort (n=135), and controls from the BRRD cohort (n=378). Exact multivariate logistic regression, of individuals of South Asian ancestry from the HCMR and BRRD cohorts, provided evidence in support of disease association for the *MYBPC3* c.1224-52G>A variant (OR: 15.90; 95% CI: 2.05 – ∞; p-value=0.003), but not the *MYBPC3*Δ25 variant (OR 1.76; 95% CI: 0.77 – 4.36; p-value=0.15) (Table 4.6). [180]

Thus, it is apparent that the *MYBPC3*Δ25 variant does not independently cause HCM. Instead, a proportion of *MYBPC3*Δ25 variant carriers possess *MYBPC3* c.1224-52G>A, which does cause HCM but has not previously been recognised, and consequently the *MYBPC3*Δ25 variant has been misrepresented as a disease-causing variant.

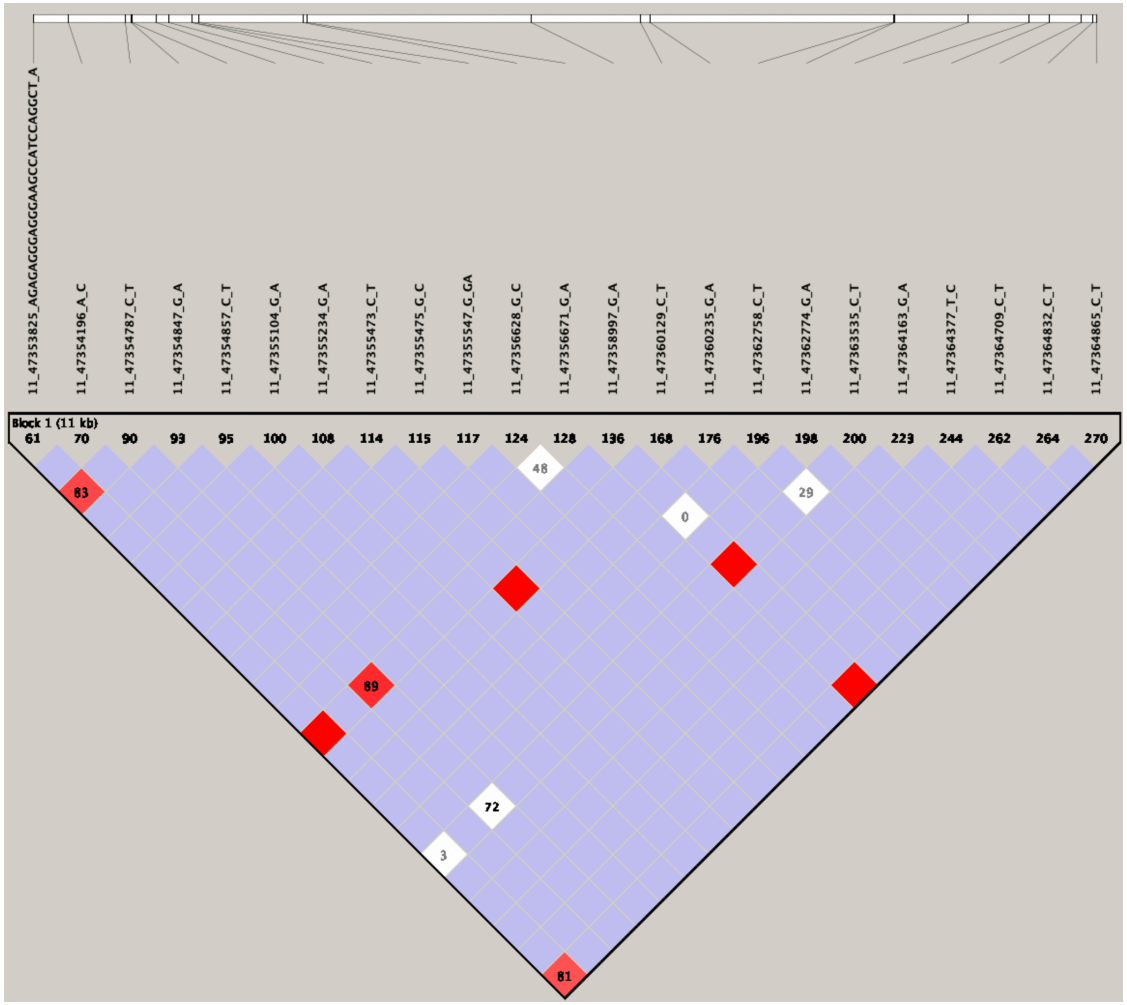


Figure 4.4: LD plot for *MYBPC3*. Generated using Haploview, red squares indicate high LD (LD) between markers. Between marker 61 (*MYBPC3*Δ25) and marker 270 (*MYBPC3* c.1224-52G>A) there is evidence of high LD ($D' = 0.81$ and $r2 = 0.22$).

A

		MYBPC3 c.1224-52G>A		
		-52/-52	**-52/+**	**+/+**
MYBPC3$^{\Delta 25}$	**Del/del**	0	0	0
	Del/+	0	5	12
	+/+	0	1	116

B

Haplotypes	**Haplotype frequencies**	
Haplotypes MYBPC3$^{\Delta 25}$ - MYBPC3 c.1224-52G>A	**LD model**	**Equilibrium model**
Del/-52	0.018	0.001
+/-52	0.004	0.021
Del/+	0.045	0.062
+/+	0.933	0.916

C

	Predicted haplotype counts	
	MYBPC3 c.1224-52G>A	**+**
MYBPC3$^{\Delta 25}$	4.8	12.1
+	1.1	250.0

Table 4.4: Haplotype analysis between *MYBPC3* c.1224-52G>A relative to *MYBPC3*Δ25 Panel A: Observed occurrences of *MYBPC3* c.1224-52G>A relative to *MYBPC3*Δ25 within the HCMR South Asian cohort (n=134); Panel B: Reported haplotype frequencies under a LD model; Panel C: Predicted haplotype counts given observed counts and LD estimates between genetic markers. + indicates the presence of the wild-type (common, assumed ancestral) allele.

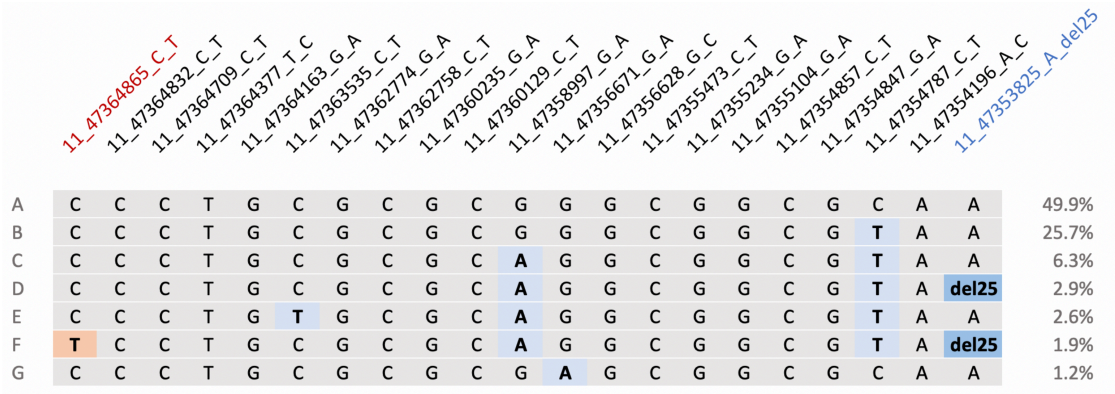


Figure 4.5: Haplotype structure across MYBPC3. Each horizontal line (denoted A-G) represents a unique haplotype predicted across MYBPC3 within the South Asian population derived from the HCMR cohort (n=134). Figure generated from data provided by Haploview. Grey indicates presence of the presumed ancestral allele, based on allele frequency. Blue shading indicates the presence of an alternate allele. The *MYBPC3*Δ25 allele is emphasised using a darker shade of blue. Red shading represents the presence of the *MYBPC3* c.1224-52G>A allele. Haplotype A is composed entirely of reference alleles and is present in 49.9% of the cohort. *MYBPC3*Δ25 is present on haplotypes D and F. Haplotype F also includes *MYBPC3* c.1224-52G>A.

	Cases				Controls					
	OMGL	HCMR	Total cases	P-value†	gnomAD genomes	Total gnomAD	TOPMED	Total Controls‡	OR (95% CI)	Fisher's P-value
MYBPC3^{-52} minor allele frequency										
Global	0.00580 (32/2,757)	0.004 (23/2,636)	0.005 (55/5,393)	0.359	3.2×10^{-5} (1/15,667)	3.2×10^{-5} (1/15,667)	- (0/62,784)	6.57×10^{-6} (1/76,048)	780 (135 – 16,384)	5.77×10^{-64}
European (NFE)	NA	0.004 (17/2,074)	0.004 (17/2,074)	-	- (0/7,696)	- (0/7,696)	No ancestry data	- (0/7,696)	∞ (15.4 – ∞)	3.43×10^{-12}
South Asian	NA	0.022 (6/134)	0.022 (6/134)	-	- (0/0)	- (0/0)	No ancestry data	- (0/0)	-	-
MYBPC3$^{\Delta 25}$ minor allele frequency										
Global	0.00363 (20/2,757)	0.003 (18/2,636)	0.004 (38/5,393)	0.98	9.56×10^{-5} (3/15,695)	0.00357 (981/139,954)	2.39×10^{-5} (3/62,784)	0.00254 (984/197,114)	1.41 (0.99 – 1.96)	0.040
European (NFE)	NA	- (0/2,074)	- (0/2,074)	-	- (0/7,708)	7.82×10^{-6} (1/63,902)	No ancestry data	7.82×10^{-6} (1/63,902)	-	-
South Asian	NA	0.063 (17/134)	0.063 (17/134)	-	- (0/0)	0.0321 (981/15,296)	No ancestry data	0.0321 (981/15,296)	1.98 (1.11 – 3.50)	0.015

Table 4.5: Association analysis of *MYBPC3*Δ25 using available case-series and reference controls from population cohorts
Case series data (n=5,393) derived from 2 large HCM cohorts (OMGL and HCMR). Reference controls derived from gnomAD and TOPMed

	***MYBPC3*$^{\Delta 25}$ carrier**		***MYBPC3*$^{\Delta 25}$ non-carrier**	
	***MYBPC3* c.1224-52G>A carrier**	***MYBPC3* c.1224-52G>A non-carrier**	***MYBPC3* c.1224-52G>A carrier**	***MYBPC3* c.1224-52G>A non-carrier**
HCMR cases	5	12	1	116
BRRD controls	0	21	0	357

Table 4.6: 2-by-2-by-2 contingency table comparing *MYBPC3*Δ25 and *MYBPC3* c.1224-52G>A in individuals of South Asian ancestry Case data derived from HCMR and control data derived from BRRD.

4.4 Discussion and limitations

In this Chapter I have conducted analysis to further evaluate the contribution of oligogenicity towards HCM using two separate approaches, specifically: 1) a systematic individual-level evaluation for multiple low-frequency variants amongst 8 core sarcomere genes; and 2) analysis driven by a candidate-variant, *MYBPC3*Δ25, thought to be a key component in the oligogenic architecture of HCM.

Unfortunately, systematic individual-level evaluation for multiple low-frequency variants amongst 8 core sarcomere genes was not adequately powered to systematically evaluate the oligogenic landscape of HCM. Nevertheless, analysis performed using *MYBPC3*Δ25 provided an insight into this proposed oligogenic architecture of HCM. Specifically, detection of a synthetic association between *MYBPC3*Δ25 and a frequently observed pathogenic variant (*MYBPC3* c.1224-52G>A; OR: 15.90; 95% CI: 2.05 – ∞; p-value=0.003), reaffirmed that HCM is a monogenic disorder, and refutes previous assumptions that *MYBPC3*Δ25 is a cornerstone upon which models of HCM oligogenicity could be based.

Whilst this is not evidence to definitively reject the hypothesis that a proportion of HCM is attributable to oligogenicity, it suggests future studies designed to systematically evaluate for oligogenicity will need to have a substantially larger case cohort to ensure adequate power if multiple rare (MAF <0.01) variants are to be pursued (Figure 4.1). In addition to a large case cohort, the study of oligogenicity is contingent on available individual-level (rather than summary-level) sequence data and genome-wide genotype data to facilitate stringent quality control procedures.

As a result there appear to be two broad approaches to further study oligogenicity in the near term, either by: 1) performing family-based genome or exome sequencing with functional evaluation of candidate variants; or 2) conducting case-control genome-wide association analyses, with consideration for common rather than rare variants, and subsequently assessing genetic risk scores across individuals.

Family based approaches have previously been demonstrated with success for left ventricular non-compaction cardiomyopathy.[98] Families with multiple affected individuals, without evidence to suggest the presence of a disease-causing

variant, would theoretically be enriched for variants that may contribute towards an oligogenic mechanism. Proving oligogenicity in such families is contingent on both *in silico* prediction tools, that identify and subsequently prioritise candidate variants, but also downstream functional analysis.

In the short-term, a more realistic goal might be to perform a genome-wide association analysis for common (MAF > 0.01) variants in HCM to evaluate their contribution towards the genetic architecture of HCM. Inferring oligogenicity via genome-wide association analyses is contingent on the underlying power of the experiment. However, anticipating *a priori* experimental power can be challenging as the true odds ratio cannot always be anticipated. For most GWAS performed across common complex disease, only modest associations (OR: 1.0-1.2) have been realised following the assembly of large cohorts, typically in the tens, if not hundreds, of thousands of participants, or more.[257]. It could therefore be assumed that large cohorts would be necessary to identify genetic associations for HCM, but this may not be true. Certainly, GWAS have previously yielded association signals with relatively small case-control cohorts. Perhaps the best known example is age-related macular degeneration, where the evaluation of 96 cases and 50 controls revealed a common intronic variant, rs380390, of large effect (OR:7.4 [95% CI: 2.90 - 19.0]; p-value = 4.1×10^{8}) in the complement factor F gene. [258] Similarly, the evaluation of several thousand cases and controls in the Wellcome Trust Case Control Consortia studies for type 1 diabetes, revealed loci in human leukocyte antigen (HLA) regions that were of relatively large effect size. [259] Whether the same is true for HCM is relatively unknown; although, one prior example does suggest the true odds ratios for common variants associated with HCM may be larger than traditionally observed for common complex traits.[118] If proven true, it would be tractable to assess oligogencity via the construction of polygenic risk score for common variants.

Projecting further into the future, systematic genome-wide analyses might facilitate the detection of rare variants of moderate effect, similar to what has been reported for non-syndromic coarctation of the aorta, where *MYH6* p.Arg721Trp (MAF in gnomAD v2.1.1 = 3.19×10^{-5}) was found to confer an OR of 44.2 [95%

CI: 20.5 – 95.5; p-value = 5×10^{-22}] in Iceland.[260] However, to achieve such a result requires access to high quality genome sequencing for both cases and controls due to concerns regarding the validity of rare variants called using genotyping arrays.[261] Presently, genome sequencing has been performed for fewer than 1,000 HCM cases from the BRRD and GeL cohorts.

5 Penetrance

Contents

5.1 Background

HCM demonstrates incomplete penetrance and variable expressivity. Penetrance relates to the probability that an individual with a disease-causing variant will

develop the associated disease. Expressivity refers to the phenotypic spectrum, specifically the severity of the phenotype, associated with a disease-causing variant.[262] Classically the penetrance of disease-causing variants has been used to assist with the genetic counselling of affected families, although it is acknowledged that the accuracy of these estimates can be fallible. Consequently, penetrance tends to be communicated to patients in relatively broad terms; it is known that HCM demonstrates age-associated penetrance, and it is understood that that some *MYH7* missense variants are more penetrant than truncating variants in *MYBPC3*.[179, 263] Similarly, in the research environment, whilst the concept of penetrance has been integrated into methodological considerations, such as when maximum credible allele frequency thresholds for disease-causing variants are calculated, the values used (i.e. 50% in HCM) tend not to be empirically derived and largely serve as a placeholder, representing a "known unknown", until more refined estimates are available.[94]

Generating accurate penetrance estimates has proved challenging in HCM. Prior efforts relied on individuals from small family-based case series, hypothesised to be enriched for genetic and/or environmental susceptibility towards disease. However, this inherent ascertainment bias resulted in artificial inflation of the overall penetrance estimates quoted. Statistical methodologies were developed to try to adjust for this possible ascertainment bias.[264]

The advent of genome-wide sequencing across large, unselected cohorts provided an alternative approach to infer penetrance estimates, but these approaches also have methodological inadequacies and may not yet be tractable for many rare diseases.[261] Specifically, population-based studies often rely on a cross-sectional, retrospective case-note review, or Hospital Episode Statistics (HES), to assign affected status, rather than prospectively examining for a phenotype, which can inflate the false negative rate in conditions where a phenotype may remain concealed unless explicitly searched for.[265] Furthermore, one would hypothesise that given the high rates of genetic and allelic heterogeneity present in conditions such as HCM, precise penetrance estimates will only be formulated when extremely large

cohorts exist, and even then, precise penetrance estimates may only be available on a small number of frequently observed variants.

A further difference between the family-based and population-based methodologies is that population-based penetrance estimates incorporate individuals in possession of secondary findings (SFs). SFs are disease-causing variants found in individuals without a personal or family-history of the corresponding disease.[266]

Clinically, how best to manage individuals with SF remains uncertain, given that the ACMG and European Society of Human Genetics (ESHG) voice opposing views. Further evidence regarding the prevalence and penetrance of SFs is required, both at an individual level, but also to assist with guidance issued through health service delivery policy recommendations. Proponents of the ACMG model argue that detection and disclosure of SFs can be beneficial for preventative medicine.[266, 267] Central to the ACMG SF recommendations are a list of 59 genes, predominated by ICC and inherited cancer syndrome genes, that are deemed to be clinically actionable and recommended to be included on all clinical sequencing studies, with findings disclosed to individuals (Table 5.1).[266, 267]

The ACMG selected these 59 genes given causal associations with conditions that demonstrate long latency periods, and for which screening and surveillance programmes can reduce morbidity and mortality. However, this differs from the perspective advocated by the ESHG, which recommends the active avoidance of SF generation due to concerns regarding the clinical utility, and the ethical implications of using a diagnostic test for the purposes of population screening.[268] Consequently further evaluation is anticipated before policy recommendations are provided to the NHS Genomic Medicine Service.[269] Therefore, the key objectives for this Chapter are to:

1. Explore the feasibility of using the UKBB to estimate penetrance estimates for HCM.

2. Generate penetrance estimates for disease-causing variants in HCM using resources from the UKBB, OMGL, HCMR, gnomAD and TOPMed, as outlined in Appendix A.1.

3. Evaluate the prevalence and associated expressivity of HCM SFs.

Disease	Gene
Inherited cardiac condition	
Arrhythmogenic right ventricular cardiomyopathy	*TMEM43, DSP, PKP2, DSG2, DSC2*
Brugada syndrome	*SCN5A*
Catecholaminergic polymorphic ventricular tachycardia	*RYR2*
Dilated cardiomyopathy	*LMNA, MYBPC3*
Fabry's disease	*GLA*
Hypertrophic cardiomyopathy	*MYBPC3, MYH7, TPM1, PRKAG2, TNNI3, MYL3, MYL2, ACTC1*
Left ventricular noncompaction	*TNNT2*
Long QT syndrome	*KCNQ1, KCNH2, SCN5A*
Other cardiovascular	
Familial hypercholesterolemia	*APOB, LDLR, PCSK9*
Familial thoracic aortic aneurysm	*MYH11, ACTA2*
Marfan's syndrome	*FBN1, TGFBR1*
Loeys-Dietz syndrome	*TGFBR1, TGFBR2, SMAD3*
Inherited cancer syndrome	
Adenomatous polyposis coli	*APC*
Breast-ovarian cancer	*BRCA1, BRCA2*
Familial medullary thyroid carcinoma	*RET*
Juvenile polyposis syndrome	*BMPR1A, SMAD4*
Li-Fraumeni syndrome	*TP53*
Lynch syndrome	*MLH1, MSH2, MSH6, PMS2*
Multiple endocrine neoplasia	*MEN1, RET*
MYH-associated polyposis	*MUTYH*
Neurofibromatosis	*NF2*
Paragangliomas	*SDHD, SDHAF2, SDHC, SDHB*
Peutz-Jeghers syndrome	*STK11*
PTEN hamartoma tumor syndrome	*PTEN*
Retinoblastoma	*RB1*
Von Hippel-Lindau syndrome	*VHL*
Wilms' tumor	*WT1*
Other	
Ehlers-Danlos syndrome	*COL3A1*
Malignant hyperthermia	*RYR1, CACNA1S*
Ornithine carbamoyl transferase deficiency	*OTC*
Pilomatrixoma	*MUTYH*
Tuberous sclerosis	*TSC1, TSC2*
Wilson disease	*ATP7B*

Table 5.1: List of 59 genes deemed clinically actionable by the American College of Medical Genetics and Genomics (ACMG).

5.2 Feasibility of using the UK Biobank to estimate penetrance estimates for HCM

5.2.1 Clinical prevalence of HCM in the UK Biobank

HCM is a disease with a typical onset in the second and third decades of life, well below the average age of UK Biobank participants (56.6 years old ($\pm$ 8.0)), suggesting that the majority of individuals likely to develop HCM will already express the phenotype. It is assumed that 0.05% of the population will have phenotypic features of HCM and carriage of a likely-pathogenic/pathogenic variant. This is based on the phenotypic prevalence of HCM being 1:500 and the genotypic prevalence for disease-causing variants in individuals with a HCM phenotype being $\sim$1/4. Whilst a population prevalence of 1:500 may be true, analysis from $\sim$4 million individuals in England, between 1997 and 2010 from primary care, hospital and mortality records via CALIBER (https://www.ucl.ac.uk/health-informatics/caliber), suggests a clinical diagnosis is only reached in 1:2,500 individuals.[13, 270] Within the UK Biobank 386 individuals (0.077% [95% CI: 0.070 - 0.085%]) have a recorded diagnosis, either self-reported or through HES data, indicating a higher than expected rate of HCM diagnoses (observed:0.077% vs. expected:0.040%; p-value $< 2.2\times10^{-16}$). It would be anticipated that 1,005 individuals (i.e. the total UKBB population (n=502,543) multiplied by the prevalence of HCM (1:500)) in the UKBB will have phenotypic evidence of HCM, which suggests 38.4% [95% CI: 35.4 - 41.5] of individuals with HCM in the UKBB have received a clinical diagnosis. However, this is assuming individuals provided with a HCM diagnosis also demonstrate phenotypic features of HCM and have not been misdiagnosed. To further explore the prevalence and concordance of HCM diagnoses, phenotypically defined HCM derived from CMR imaging could be compared with self-reported/HES data (individuals with an ICD10 classification of I42.1 (obstructive HCM) and I422 (other HCM)) collected for individuals who underwent CMR imaging as part of the UK Biobank (6.60%; n=33,146/502,543). However, to perform such analysis

is reliant on the generation of maximum wall thickness measurements, which have not yet been publicly released to the UKBB community.

5.2.2 Prevalence of pathogenic/likely pathogenic variants

Of the 386 individuals deemed to have HCM in the UKBB, 5.18% (n=20/386) were amongst the 49,960 individuals included in the first tranche of UKBB exome sequencing.[131] To further evaluate UKBB exome sequence data, PLINK files mapped against GRCh38 were downloaded for UKBB application 11223. Disease-specific regions of interest, across 8 core sarcomere genes (*MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3, ACTC1* and *TPM1*) were extracted for downstream analysis. Genomic co-ordinates for GRCh37 were annotated using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

It was surprising to find no likely pathogenic or pathogenic variants, across 8 core sarcomere genes, present in the 20 individuals with HCM who underwent exome sequencing. However, there was evidence of SFs, across the 8 core sarcomere genes; 109 of the 49,960 individuals who underwent exome sequencing possessed either a pathogenic (n= 69, 0.14%), or likely pathogenic (n=40, 0.09%) variant without a HCM diagnosis. Variants of uncertain significance were detected in an additional 461 individuals (0.92%). It is noteworthy that *MYBPC3* p.Arg502Trp was the most frequently observed SF, observed in 23.9% ([95% CI: 16.8 – 32.7]; n=26/109) of individuals with a SF. It was surprising that no pathogenic or likely pathogenic SFs were detected in *MYH7*, as *MYH7* is a predominant source of disease-causing variants in HCM (Chapter 3).

To further evaluate the absence of SFs in *MYH7* observed in the UKBB, non-overlapping reference controls from gnomAD and TOPMed (n=198,517) were combined and assessed across 8 core sarcomere genes. Across gnomAD and TOPMed, the prevalence of pathogenic or likely pathogenic variants was 0.36% [95% CI: 0.33 – 0.39], with 127 different pathogenic or likely pathogenic variants detected in 714 individuals (Table 5.2). *MYH7* (n=64/127) and *MYBPC3* (n=49/127) harboured the largest proportion of disease-causing variants. This calculation has assumed

that each individual from the gnomAD and TOPMed cohort possesses a maximum of one disease-causing allele (i.e. no homozygous carriers and no digenic or double heterozygous carriers). This assumption is obligatory given a lack of individual-level data and understanding that HCM is an autosomal dominant condition with <1% of affected individuals demonstrating multiple pathogenic/likely pathogenic variants.[176] Further, it is assumed that gnomAD and TOPMed are not enriched for individuals diagnosed with HCM. If an adjustment is made to account for the possible presence of HCM individuals (from 198,517 across gnomAD/TOPMed it is theoretically possible 397 individuals could have HCM (i.e. 1:500), and 107 may yield a disease-causing variant) this results in a slightly more conservative prevalence estimate of 0.31% [95% CI: 0.28 – 0.33].

Consequently, given that disease-causing *MYH7* variants are present in gnomAD/TOPMed, finding no SFs in *MYH7* in the UKBB cohort is most likely representative of a technical artefact. This may relate to the wider public concern pertaining to the UKBB functional equivalent (FE) exome sequencing dataset, attributable to an issue with duplicate read marking and the underreporting of rare variants.[271–273] It is not possible to accurately estimate the prevalence of SFs in the UKBB until the scheduled public release of 150k UKBB exomes.[271] The presence of false-negatives in the UKBB FE data may explain why, relative to gnomAD/TOPMed, following the adjustment for possible inclusion of undisclosed HCM cases, there is a depletion of SFs in the UKBB (i.e. UKBB: 109/49,960 (0.22%) vs. gnomAD/TOPMed: 607/198,120 (0.31%); p-value = 0.001).

Data from gnomAD/TOPMed indicates that the prevalence of SFs in the general population is ∼0.31%, but appreciating the phenotypic prevalence of HCM in SF carriers is not yet known. To support the provision of necessary healthcare services, approximating the expressivity of HCM in SF carriers is necessary, assuming the classification of likely-pathogenic and pathogenic variants is accurate.

Gene	HGVS.c	HGVS.p	ACMG classification	Allele Count	Cohort
MYBPC3	c.1504C>T	p.Arg502Trp	5	46	UKBB 26; GT 20
MYBPC3	c.1624G>C	p.Glu542Gln	5	22	UKBB 10; GT 12
MYBPC3	c.3330+5G>C	-	5	20	UKBB 2; GT 18
MYBPC3	c.772G>A	p.Glu258Lys	5	14	UKBB 2; GT 12
MYBPC3	c.624G>C	p.Gln208His	4	82	UKBB 3; GT 79
MYH7	c.1322C>T	p.Thr441Met	4	79	GT 79
MYH7	c.3981C>A	p.Asn1327Lys	4	42	GT 42
MYBPC3	c.787G>A	p.Gly263Arg	4	31	UKBB 2; GT 29
MYBPC3	c.2504G>T	p.Arg835Leu	4	22	GT 22
MYH7	c.3158G>A	p.Arg1053Gln	4	18	GT 18
MYBPC3	c.1591G>A	p.Gly531Arg	4	15	GT 15
MYH7	c.2606G>A	p.Arg869His	4	15	GT 15
MYBPC3	c.3065G>C	p.Arg1022Pro	4	14	UKBB 6; GT 8
MYBPC3	c.2459G>A	p.Arg820Gln	4	14	UKBB 6; GT 8
MYL2	c.403-1G>C	-	4	13	GT 13

Table 5.2: Pathogenic or likely pathogenic variants associated with HCM detected in reference controls with a combined allele count greater than 12
GT: gnomAD/TOPMed; UKBB: UK Biobank

5.2.3 Penetrance estimates from the UK Biobank

UKBB participants underwent genotyping, and will eventually also undergo sequencing, irrespective of clinical features. Consequently, the UKBB represents a platform that has adopted a "genotype-first" approach, circumventing the ascertainment bias inherent to previous studies. This provides a theoretical opportunity to generate per-variant penetrance estimates for disease-causing HCM variants. However, such approaches in HCM are currently constrained by the low prevalence of HCM in the UKBB (0.08%, n= 386/502,543), substantial allelic and genetic heterogeneity and possible sequencing and phenotyping errors (discussed above).

Assuming HCM cases have been correctly diagnosed in the UKBB, an assumption that will need to be reviewed following the public release of maximum wall thickness measurements derived from cardiac magnetic resonance imaging, an alternative approach would be to use array-based genotypes. However, array-based genotyping was designed to assay common genetic variation and tends to perform poorly when re-purposed to assess rare genetic variation due to an inherent reliance on assigning genotypes based on genotype clusters. Algorithms designed to differentiate

true genotype signals from experimental noise struggle when very few individuals are carriers of the alternative allele.[156] Consequently, for many rare variants, genotyping-arrays provide inaccurate results and may not be suitable for the purposes of penetrance calculations.[261] However, following the systematic evaluation of rare variants captured by the UKBB array, through visual inspection of genotype cluster plots (performed by Wright et al. (2019)), 16 rare variants associated with HCM were deemed to demonstrate satisfactory quality.[261] Extracting genotype counts from the UKBB array data, for HCM cases and controls, allows odds ratios to be generated using Fishers exact test. *MYBPC3* p.R502W is the most frequently detected variant in UKBB HCM cases (1.61% [95% CI: 0.74-3.47%]), comparable to observations from the OMGL-HCMR case series (Table 5.3). Of the variants available, *MYH7* p.D906G appears to be the most penetrant (0.094 [95% CI:0.011-0.76]), but the confidence intervals associated with all these variants are wide, limiting clinical utility. However, ultimately uncertainty regarding the case definition of HCM is what presently undermines these estimates. Until a study formally assesses correlation between UKBB assigned HES codes and phenotypic appearances on cardiac magnetic resonance imaging for individuals presumed to have HCM, no penetrance estimates should inform clinical decision making.

Gene	c.HGVS	p.HGVS	ACMG	Cases		Controls		Odds ratio [95% CI]	Penetrance [95% CI]
				AC	AN	AC	AN		
MYH7	c.A2717G	p.D906G	5	1	742	28	974,848	46.92 [6.24-352.89]	0.09 [0.01-0.76]
MYBPC3	c.C1504T	p.R502W	5	6	732	204	974,748	39.17 [16.31-94.06]	0.08 [0.03-0.19]
MYL2	c.A401C	p.E134A	3	2	740	420	973,302	6.26 [1.27-30.96]	0.01 [0-0.05]
TNNT2	c.C814T	p.R272C	4	2	740	626	973,644	4.2 [0.78-22.75]	0.01 [0-0.03]
MYL2	c.C141A	p.N47K	3	1	742	771	974,206	1.7 [0.14-20.18]	0 [0-0.02]
MYBPC3	c.G13C	p.G5R	3	1	742	1,167	972,210	1.12 [0.08-16.67]	0 [0-0.01]
MYH7	c.T2945C	p.M982T	1	1	742	2,234	970,650	0.59 [0.02-14.76]	0 [0-0.01]
TNNT2	c.C838T	p.R280C	3	0	744	13	975,068	50.41 [2.95-862.7]	0 [0-1]
TNNT2	c.G714T	p.E238D	3	0	744	218	974,576	3 [0.15-60.24]	0 [0-0.05]
MYBPC3	c.C3137T	p.T1046M	3	0	742	32	975,370	20.54 [1.21-347.54]	0 [0-0.44]
MYBPC3	c.C2618A	p.P873H	3	0	744	103	974,740	6.36 [0.35-114.09]	0 [0-0.12]
MYBPC3	c.G2497A	p.A833T	2	0	744	1,993	970,532	0.33 [0-26.86]	0 [0-0.01]
MYBPC3	c.G223A	p.D75N	3	0	744	36	975,310	18.21 [1.07-308.65]	0 [0-0.39]
MYBPC3	c.26-2A>G	.	3	0	742	41	975,292	16.03 [0.94-272.49]	0 [0-0.33]
MYH7	c.C4130T	p.T1377M	4	0	744	1,138	972,058	0.57 [0.01-25.49]	0 [0-0.01]
TNNI3	c.407G>A	p.R136Q	3	0	744	9	974,962	72.8 [4.17-1270]	0 [0-1]

Table 5.3: Effect size estimates for variants deemed to be of adequate quality by Wright et al. (2019)[261] RefSeq Transctipts used for each gene: *TNNT2*:NM_001001432; *MYBPC3*:NM_000256; *MYL2*:NM_000432; *MYH7*:NM_000257; *TNNI3*:NM_000363. Abbreviations: AC=allele count; AN=allele number. Penetrance calculated using formula outlined by Minikel et al. (2016) as described in chapter 3. [256, 261]

5.3 Generate penetrance estimates for disease-causing variants in HCM

5.3.1 Case-control allele frequency comparison

A case-control study design, comparing allele frequencies for variants assigned uncertain significance, likely pathogenic and pathogenic status, from across 12 cardiomyopathy associated genes (*MYBPC3, MYH7, TNNT2, TNNI3, MYL2, MYL3, ACTC1, TPM1, GLA, PRKAG2, LAMP2* and *TTR*), was undertaken to retrospectively calculate penetrance estimates. HCM cases were derived from the OMGL and HCMR cohorts, and gnomAD exomes acted as reference controls (Figure 5.1). Plotting variants, stratified by ACMG classification, revealed two main outliers on the gnomAD allele frequency axis (Figure 5.1A), specifically: *MYBPC3*Δ25 and *TTR* p.Val142Ile. The biological relevance of *MYBPC3*Δ25 to HCM is explored in Chapter 4. Pathogenic variants in *TTR* are associated with amyloidosis, a life-threatening, multi-system disorder that is typically associated with a restrictive amyloid cardiomyopathy. Consequently, *TTR* is often screened on HCM gene panels as a phenocopy gene. The *TTR* p.Val142Ile variant is one of the most commonly observed variants in *TTR* amyloidosis, particularly in individuals of African ancestry where the carrier frequency has been reported to be as high as $\sim$3.5%, and is well-recognised as having variable penetrance and expressivity.[274] As *TTR* p.Val142Ile does not principally pertain to HCM, but is instead the cause of an established HCM phenocopy, further investigation into this variant will not be presented in this thesis.

5.3.2 Distribution of case-control allele frequencies stratified by ACMG classification

By limiting the allele frequency in gnomAD exomes to exclude these extreme outliers, a distinct distribution for variants from each ACMG classification grouping can be observed (Figure 5.1B). Likely pathogenic and pathogenic variants are observed less frequently in gnomAD than variants of uncertain significance and seen more often in the HCM cohort. Two likely pathogenic variants, specifically *MYBPC3* p.Gln208His and *MYH7* p.Asn1327Lys, demonstrated allele frequencies in gnomAD that exceeded

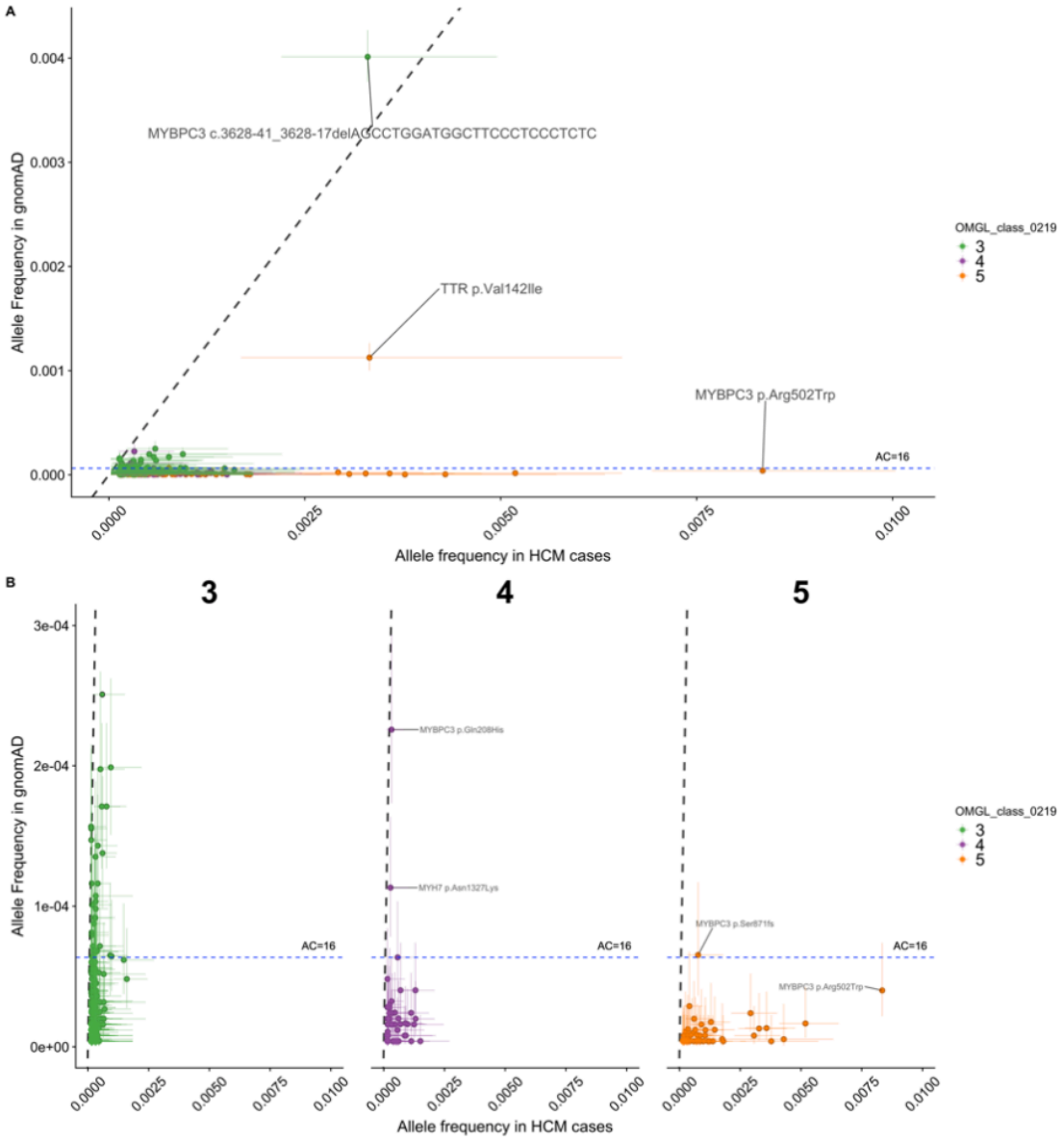


Figure 5.1: Case-control comparison of allele frequencies for disease-causing variants Variants of uncertain significance (green), likely pathogenic (purple), and pathogenic (orange) variants in HCM. Blue horizontal line represents maximum credible allele frequency threshold for monogenic HCM. Black oblique line represents line of equivalence.

those expected of a true disease causing variant (i.e. an allele count of 16 based on the size of the gnomAD cohort), based on the weak assumption that pathogenic variants have a maximum penetrance of 50%.[94] Both *MYBPC3* p.Gln208His and *MYH7* p.Asn1327Lys were classified by the OMGL team before gnomAD was available, when allele frequencies across diverse population groups were difficult to

obtain for variant classification purposes. Access to gnomAD demonstrates that both variants are polymorphisms specific to the Ashkenazi Jewish population and therefore unlikely to be pathogenic. This finding was communicated to the OMGL team so that re-classification can be formally documented, and patients contacted regarding this re-classification. Variants of uncertain significance that exceeded this threshold were provided to the OMGL team for re-review.

5.3.3 Relationship between odds ratios and variant classification

By comparing the allele frequencies of variants assigned an ACMG classification by the OMGL team between cases and controls, per-variant odds ratios were generated. This demonstrates a positive correlation with pathogenic variants on average demonstrating the largest odds ratios. However, there is considerable variation within each group (Figure 5.2).

5.3.4 Approximating penetrance of secondary findings

Allele frequencies derived from case-series data, encompassing the HCMR cohort (n=2,636) and the OMGL HCM cohort (n= 2,757), were compared with non-overlapping unscreened controls in gnomAD and TOPMed (n=198,517). 26.9% [95% CI: 25.7 – 28.1] of HCM cases were found to carry either a pathogenic (n= 1,043) or likely pathogenic (n=407) variant; 121 variants were present and demonstrated satisfactory coverage within gnomAD or TOPMed.

The use of unscreened population controls in genetic association studies investigating common disease genetics is well established. However, in rare disease genetics, where large effect alleles underpin disease aetiology, the presence of misclassified cases/controls has the potential to appreciably dilute true effect sizes, as low effect alleles for common disease are also diluted to a trivial extent.[275, 276] To account for the potential presence of misclassified cases/controls a positive likelihood ratio framework, established by Moskvina et al. (2005), was implemented to provide penetrance estimates for disease-causing variants.[276] Application of this

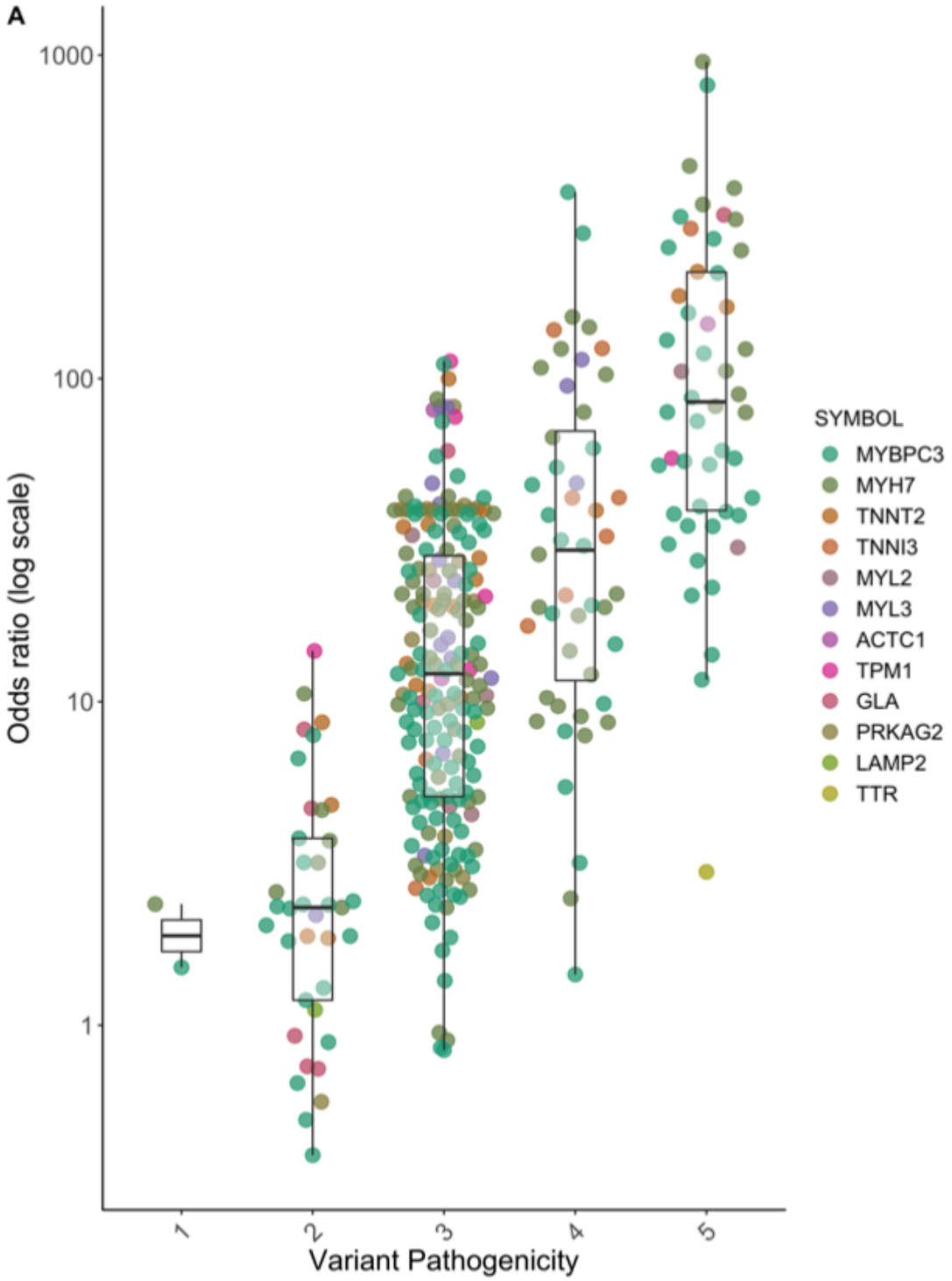


Figure 5.2: Relationship between ACMG variant pathogenicity and odds ratio across 12 genes causally associated with HCM

framework is specific to the penetrance associated with a SFs, as the prior probability implemented is reflective of the population prevalence of HCM (i.e. 1/500).

As such, per variant allele counts were represented in a 2 × 2 contingency table, as follows:

	Cases	Controls
Allele1	*a*	*b*
Allele2	*c*	*d*
	$2n_1=a+c$	$2n_2=b+d$

For cases, the relative allele frequency (p_1) is calculated using equation I. For controls, the relative allele frequency (p_u) would typically be calculated using equation II. However, to account for the potential presence of misclassified cases amongst unscreened controls, the relative allele frequency for controls (p_2) is calculated, when, $0 \leq p_2 \leq 1$, using equation III:

$$p_1 = \frac{a}{2n_1} \tag{I}$$

$$p_u = \frac{b}{2n_2} \tag{II}$$

$$p_2 = \frac{p_u - k_p \times p_1}{[1 - k_p]} \tag{III}$$

k_p represents the population prevalence of disease (i.e. 1/500).

Using p_1 and p_2, predicted odds ratios were calculated (equation IV), alongside sensitivity (equation V) and specificity (equation VI):

$$OR = \frac{p_1[1 - p_2]}{p_2[1 - p_1]} \tag{IV}$$

$$sensitivity = \frac{p_1}{p_1 + [1 - p_1]} \tag{V}$$

$$specificity = \frac{1 - p_2}{[1 - p_2] + p_2} \tag{VI}$$

The positive likelihood ratio (equation VII), post-test odds (equation VIII) and post-test probability (equation IX) were then calculated:

$$LR(+) = \frac{sensitivity}{1 - specificity} \tag{VII}$$

$$PTO = \frac{LR(+) \times k_p}{1 - k_p} \tag{VIII}$$

$$PTP = \frac{PTO}{1 + PTO} \tag{IX}$$

The variance of the natural logarithm of the odds ratio was calculated using (equation X):

$$\sigma^2 = \frac{1}{2n_1} \cdot \left[\frac{1}{p_1[1 - p_1]} + \frac{1}{r \times p_2[1 - p_2]} \right] \tag{X}$$

Where $r = n2/n1$ (i.e. the number of controls to the number of cases).

Penetrance estimates were calculated for 121 variants deemed to be pathogenic or likely pathogenic by ACMG guidelines. However, allelic heterogeneity inherent to HCM results in relatively imprecise per-variant penetrance estimates for almost all variants evaluated. Only four variants were present in more than 10 cases and 10 controls, which facilitated more precise penetrance estimates, but substantial heterogeneity was present (Table 5.4). The extended list of penetrance estimates is presented in appendix A.5. The most commonly observed pathogenic variant in HCM, *MYBPC3* p.Arg502Trp (n=97/5,393), yields an estimated penetrance of 36% [95% CI: 24 – 50%] (Table 5.4). Three variants assigned likely pathogenic status, specifically *MYH7* p.Arg1053Gln, *MYBPC3* p.Gly263Arg, and *MYBPC3* p.Gln208His demonstrate an odds ratio that spans 1.0, which suggests they may require re-evaluation (Appendix A.5).

Variant	Cases		Controls		Odds ratio [95% CI]	Penetrance [95% CI]	ACMG
	AC	AN	AC	AN			
MYBPC3; c.1504C>T, p.Arg502Trp	97	10,689	20	394,868	277 (154 – 491)	0.36 (0.24 – 0.50)	5
MYBPC3; c.772G>A, p.Glu258Lys	63	10,721	12	387,528	304 (142 – 645)	0.38 (0.22 – 0.56)	5
MYBPC3; c.1624G>C, p.Glu542Gln	28	10,758	12	377,066	97.5 (47.3 – 200)	0.16 (0.09 – 0.29)	5
MYBPC3; c.3330+5G>C	15	10,771	18	373,892	30.6 (15.3 – 61.3)	0.06 (0.03 – 0.11)	5

Table 5.4: Penetrance estimates for secondary findings Variants selected based on the presence of at least 10 alleles in both cases and controls.

5.4 Evaluate the phenotype of secondary finding carriers in HCM

5.4.1 Rationale for evaluating SFs in ICC

There is a need for robust evidence to inform interpretation and disclosure of SFs. Evaluation of SFs in inherited cardiac conditions (ICC), encompassing inherited cardiomyopathies and long QT syndrome, is perhaps the most tractable disease entity through which evidence can be generated. ICC associated SFs are relatively common, with several previous studies estimating their prevalence. However, it should be recognised that the prevalence of SFs is contingent on the genes surveyed and the variant classification approach employed. The clinical impact of SFs remains poorly defined and here SFs associated with HCM are specifically considered, with respect to prevalence, penetrance and expressivity.

5.4.2 Phenotypic evaluation of secondary finding carriers

Studies aiming to quantify the penetrance of secondary findings are often reliant on electronic medical records or the retrospective review of investigations arranged for other indications, but in the context of inherited cardiac conditions such practices may underestimate the phenotypic prevalence of disease. For example, a study performed using thirty-thousand participants in the Geisinger Health System suggested SFs with a pLoF effect in established ARVC genes (*PKP2, DSG2, DSC2, DSP, JUP, TMEM43*, or *TGFβ3*) were not associated with an ARVC phenotype.[265] This conclusion was reached following the evaluation of 14 individuals with a pLOF variant associated with ARVC, of whom only one had minor diagnostic criteria. However, the ARVC phenotype was evaluated using a retrospective case note review of echocardiography findings, rather than prospectively assessing the phenotype of variant carriers using the gold-standard diagnostic modality, cardiac magnetic resonance imaging (CMR). Given that the population prevalence of ARVC is between 1:1000 and 1:5000, it is unusual that

no individuals with proven ARVC were detected and raises concerns regarding the methodological approach employed.[277]

To address some of these methodological challenges the expressivity of HCM related SFs was evaluated using a prospective recall-by-genotype clinical study.

5.4.3 Evaluating expressivity of secondary findings in the BRRD

The SCARFE study was a recall-by-genotype, double blind, case-control feasibility study designed to evaluate the expressivity of secondary finding carriers, alongside qualitative exploration of the behavioural and psychosocial outcomes associated with secondary finding disclosure. Invitations were sent to individuals, enrolled in the BRRD, who provided consent to participate in follow-on studies using a two-stage recruitment process, in an attempt to protect participant autonomy.

Bioinformatics

Whilst the SCARFE study explored SFs in the context of multiple ICCs, data presented here are specific to HCM-associated SFs derived from 8 causally associated HCM genes (*ACTC1, MYL2, MYBPC3, MYL3, MYH7, TNNI3, TNNT2* and *TPM1*). Genomic intervals corresponding to the canonical transcript for each gene were retrieved from Ensembl (GRCh37) and extracted from the BRRD VCF file. Sequencing quality metrics were reviewed for variants predicted to be of either high or moderate impact, as determined by SNPEff, with a global minor allele frequency of less than 0.0001, as reported by gnomAD. In October 2017, variants were classified using the ACMG variant classification guidelines, as implemented by OMGL, enabling the identification of likely pathogenic and pathogenic variant carriers. Variants assigned pathogenic or likely pathogenic status were confirmed using Sanger sequencing by the BRRD.

Clinical assessment

The following criteria were used to screen for eligible individuals: between 18 and 80 years old; capable of providing informed consent; no evidence of an ICC diagnosis or

a comorbidity that may confound phenotype interpretation via echocardiogram (i.e. pulmonary hypertension), CMR imaging, or electrocardiogram; or, any objections from the individual's clinical care team sought before participants were made aware of this study. The London Fulham research ethics committee granted approval for the study (17/LO/1579).

Individuals previously consented to stage 2 studies within the BRRD cohort were enrolled using established recruitment-by-genotype methodology, with consideration for specific ethical issues including violation of an individual's "right not to know", and the possibility of causing distress by imparting genetic risk information.[278] To address these concerns, a two-step process was implemented and co-ordinated via the central BRRD team. First, eligible participants were sent a letter asking them to opt out if they did not wish to receive health or genetic information that did not pertain to the health condition for which they were enrolled in the BRRD. Eligible participants who opted out at this stage did not receive an invitation letter. Second, eligible participants who chose not to opt out, within four weeks of the opt out letter being sent, were sent an invitation letter and patient information leaflet, inviting them to undergo cardiology screening at a specialist ICC service in Oxford. Individuals wishing to participate in this recall-by-genotype study evaluating SFs in ICC genes liaised directly with the BRRD. The BRRD co-ordinating team then randomised variant carriers and non-variant carriers, with 1:1 allocation, based on their age and gender, before informing the Oxford investigators of the participant's details. Details communicated to the Oxford team included basic demographic details and the specific ICC under consideration. No information regarding variant carrier status was shared with the investigators prior to clinical evaluation, in an effort to limit bias.

Informed consent was taken by study investigators, experienced in managing patients with ICCs, with information regarding the specific ICC under consideration explained, clinical and psychosocial implications of a diagnosis, and implications for relatives.

Demographic details and a personal medical history, including a three-generation family history, were recorded for all participants, before cardiac phenotyping was performed. Cardiac investigations were performed in alignment with established clinical practice and tailored to each specific ICC. To limit heterogeneity between study participants, cardiac investigations were standardised, with the same team of experienced operators, following a standardised protocol, using the same equipment and settings for all participants. An experienced ICC cardiologist applied a systematic approach to interpret of each investigation whilst remaining blinded to variant carrier status. A clinical impression was formulated, and documented, before variant carrier status was disclosed to the investigators. Results of the investigations and variant carrier status were then disclosed to the participant, followed by appropriate counselling, correspondence to the participant's general practitioner, and follow-up arrangements determined.

Results

20 individuals, from a possible 7,203 individuals enrolled in the BRRD (0.28% [95% CI:0.18 - 0.43%]), possessed an HCM related SF (Table 5.5). The enrolment strategy for the BRRD, whilst directed towards the recruitment of individuals with a rare disease, did include unaffected family members.

Of the 20 individuals harbouring a HCM associated SF, 4 participated in a recruitment-by-genotype clinical study. Whilst all four variant carriers were classified as carrying pathogenic or likely pathogenic variant at the time of variant classification (January 2017), one variant, detected in two of the four variant carriers, was re-classified to a VUS, prior to phenotypic assessment in 2018. A summary of the clinical findings derived from variant carriers and non-variant carriers is presented in Table 5.6.

Of the four variant carriers, one had a positive family history of HCM, and was clinically affected; a second was clinically unaffected and had no relevant or suspicious family history. One variant, in two unrelated participants, was subsequently re-classified as being of uncertain significance.

Gene	HGVS.c	HGVS.p	Count	ACMG
MYBPC3	c.1504C>T	p.Arg502Trp	3	5
MYBPC3	c.1624G>C	p.Glu542Gln	1	5
MYBPC3	c.2373dupG	p.Trp792fs	1	5
MYBPC3	c.26-2A>G	-	2	4
MYBPC3	c.3192dupC	p.Lys1065fs	1	5
MYBPC3	c.3592dupG	p.Ala1198fs	1	5
MYBPC3	c.772G>A	p.Glu258Lys	2	5
MYBPC3	c.994G>T	p.Glu332*	1	5
MYL3	c.170C>G	p.Ala57Gly	2	4
MYH7	c.4066G>A	p.Glu1356Lys	1	4
TNNI3	c.433C>T	p.Arg145Trp	2	5
TNNI3	c.434G>A	p.Arg145Gln	1	4
TNNI3	c.484C>T	p.Arg162Trp	1	4
TNNI3	c.485G>A	p.Arg162Gln	1	4

Table 5.5: HCM associated secondary findings detected in the BRRD cohort.

Age (gender)	Echo (cm)	ECG	BP	FH	Affected	Variant
70yo (F)	MWT= 1.2 LA = 3.0 No hypertrophy. Normal contractility. Slightly enlarged atria	Small voltages	138/70	No	Unaffected	*MYBPC3* c.2373dupG, p.Trp792fs Pathogenic
46yo (M)	MWT = 1.5 LA= 3.5 Slightly hypertrophied mid septum, suspicious for HCM. Unremarkable contractility	Sinus Bradycardia Normal voltages with No criteria for LVH	113/65	Yes: FDR with HCM; No gene test	Affected	*MYBPC3* c.3592dupG p.Ala1198fs Pathogenic
66yo (F)	MWT= 1.2 LA = 3.0 Slight septal bulge, insufficient for diagnosis	Small Q waves in III	155/72	Yes: 2 FDRs died suddenly in 70s	Unaffected	*MYL3* c.170C>G p.Ala57Gly Likely pathogenic (VUS)
60yo (F)	MWT= 1.1 LA = 3.4 Normal septal thickness and contractility	No voltage criteria for LVH	122/72	No	Unaffected	*MYL3* c.170C>G p.Ala57Gly Likely pathogenic (VUS)
48yo (F)	MWT= 1.0 LA = 3.3 Normal contractility, normal cavity size and wall thickness. Borderline increase in LA volume	ST-T wave changes in V2/V3 of uncertain significance; no voltage criteria for LVH	122/79	No	Unaffected	None
46yo (M)	MWT= 1.2 LA = 4.2 MWT upper limit of normal; large vessels in keeping with athletic training, normal diastolic function	T wave inversion in III, aVF; slight QRS widening (120ms); prominent U wave, PR interval short (88ms). No obvious delta wave	146/98	No	Unaffected	None
67yo (F)	MWT= 1.2 LA = 3.6 Borderline localised Proximal hypertrophy, slightly dilated aortic root in keeping with hypertension	No evidence of prior infarct. PR interval upper limit of normal; shallow T wave inversion in V2. No voltage criteria for LVH	134/77	No	Unaffected	None
39yo (F)	MWT= 0.8 LA = 2.5 Normal	Normal	130/75	No	Unaffected	None
70yo (F)	MWT= 1.2 LA = 3.8 MWT upper limit of normal; normal contractility	Normal; no voltage criteria for LVH	154/84	No	Unaffected	None

Table 5.6: Summary of the clinical findings derived from variant carriers and non-variant carriers enrolled in SCARFE Abbreviations: BP: blood pressure; FH: family history; FDR: first degree relative; MWT: maximum wall thickness; LA: left atrial size; LVH: left ventricular hypertrophy

5.5 Discussion and limitations

The analysis performed in the chapter has attempted to evaluate the penetrance and expressivity of disease-causing variants associated with HCM.

Feasibility of using the UKBB to estimate penetrance estimates for HCM The UKBB was evaluated to assess whether genotype-first penetrance estimates could be generated for HCM. However, it is apparent that the UKBB is presently an inappropriate resource for the purposes of penetrance estimates in HCM. This relates to both inherent challenges to the validity of the HCM phenotype and genotyping quality. Further evaluation of individuals yielding a HCM diagnosis is required, using CMR images, to ensure the validity of this diagnostic code in the UKBB. Current public concerns regarding the genotype quality and variant calling issues demonstrated by the FE exome data cast doubt on the validity of any penetrance estimates generated from the UKBB.

Penetrance estimates for disease-causing variants in HCM Issues inherent to HCM, such as a relatively low population prevalence (1:500), alongside high allelic heterogeneity, serve as further challenges that may limit the precision of any future penetrance estimates in the UKBB. These issues were encountered when a positive likelihood ratio framework was implemented using a case-control study design (OMGL-HCMR cases and gnomAD/TOPMed controls), with only four variants in *MYBPC3* (p.Arg502Trp; p.Glu258Lys; p.Glu542Gln; and c.3330+5G>C) capable of generating relatively confident estimates.

The prevalence and associated expressivity of HCM SFs Genotype-first methodologies have led to an increasing number of individuals receiving information regarding SFs. Concerns regarding the false negative rate of SFs in the UKBB meant prevalence estimates could not be reliably generated. Instead the prevalence of HCM related SFs was calculated in the BRRD (0.28% [95% CI:0.18 - 0.43%]) and in gnomAD/TOPMed (0.31% [95% CI: 0.28 – 0.33]) using standardised ACMG

variant classifications. Prevalence estimates calculated in gnomAD/TOPMed made adjustments for the assumption that a proportion of the gnomAD/TOPMed population would have HCM at population prevalence (i.e. 1:500) and hadn't been removed. It is reassuring that estimates from the BRRD and gnomAD/TOPMed provide congruent approximations for the prevalence of SFs. Analyses evaluating the prevalence of SFs across 8 core sarcomere genes in the Framingham Heart Study (FHS) and Jackson Heart Study (JHS) had suggested a prevalence of 0.6% (22/3600), two-fold higher than what has been approximated using the BRRD and gnomAD/TOPMed.[279] It is possible though that as the FHS/JHS was performed in an era preceding the availability of large reference control datasets, such as gnomAD, many variants previously classified as likely-pathogenic or pathogenic have now been downgraded and no longer reflect SFs. Contemporary analysis performed in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort provides estimates for SFs in HCM genes (0.13% [95%CI:0.05-0.29%]) that overlaps with those derived from BRRD and gnomAD/TOPMed.[280]

A recruitment-by-genotype (RBG) study, aimed to evaluate the expressivity of SFs associated with HCM, was conducted. This RBG study was not designed to evaluate quantitative differences between individuals with SFs, but instead provided preliminary data highlighting the variable expressivity present across SF carriers. One individual, 70 years old, was found to harbour a disease-causing variant (*MYBPC3* p.Trp792fs) without any evidence of either a personal or family history of HCM and may be an example of genetic resilience.[281]

Overall, the prevalence and inherent allelic heterogeneity of HCM represents a substantial challenge for the evaluation of the penetrance and expressivity of SFs.

6

Common variant contributions

Contents

6.1 Background

For common, genetically complex conditions such as type two diabetes and hypertension, genome wide association studies (GWAS) have delivered insights into disease biology.[109, 238, 282] Beyond the delineation of novel disease biology in common complex disease, GWAS have been adopted to evaluate polygenic contributions across rare diseases, with emerging evidence that common genetic variants influence both the penetrance and expressivity of rare monogenic diseases.[283–286]

There is reason to believe that a substantial proportion of HCM may be influenced by common variants, and further evaluation is warranted. HCM represents a relatively common rare disease for which the majority of individuals remain without a genetic diagnosis, yet there is evidence of familial recurrence. Whilst this could be attributable to a shared environmental exposure, it could also be hypothesised to be attributable to genetic susceptibility to disease. When a rare variant hypothesis has been pursued for the genetic aetiology of sarcomere negative HCM, either through application of a monogenic or oligogenic model of disease, the insights that have been generated only account for 1% of disease (see Chapter 1). Polygenic contributions remain relatively unknown, with the only published evidence derived from a GWAS study performed in 2012, where *FHOD3* (lead SNP rs516514, a common intronic variant with a minor allele frequency of 0.48) demonstrated suggestive evidence for association with HCM (OR: 2.45; [95% CI: 1.76–3.41]; p-value=1.25×10^{-7}).[118] Although *FHOD3* demonstrates a role in both actin filament formation and sarcomere development, rs516514 could plausibly represent a

false positive, for two reasons. First, carriers of the rs516514 risk allele demonstrated no difference in their cardiac morphology, compared with non-carriers. Second, the discovery effort may be underpowered, as only 174 HCM cases were enrolled.[118] Whilst it is possible to achieve genome-wide significance through a case-control study design of under 1,000 cases, this is generally only observed in a pharmacogenomic setting, where a lack of evolutionary response to an exogeneous agent is likely responsible for large effect sizes from common variants.[287]

It has subsequently been reported that *FHOD3* is a disease-causing gene for HCM, following the identification of affected families with rare, co-segregating variants in *FHOD3*. [65] Nevertheless, GWAS replication is still required for the *FHOD3* locus.

Therefore, the key objectives for this Chapter are to:

1. Evaluate the common variant contribution towards HCM
2. Evaluate the genetic architecture of sarcomere positive and sarcomere negative HCM.
3. Evaluate an individual's risk of developing HCM through the aggregate burden of common genetic variants

6.2 Evaluate the common variant contribution towards HCM

A flow diagram outlining the analytical plan for this Chapter is reported in Figure 6.1.

6.2.1 Case-control GWAS of HCM

Three multi-ancestry case-control studies were conducted to evaluate the common variant contribution towards the genetic architecture of HCM. Case-control studies included: 1) HCMR cases (n=2,541) vs. UKBB controls (n=40,283); 2) BRRD: 239 cases vs. 7,203 controls and 3) UKBB: 326 cases vs. 5,767 controls.

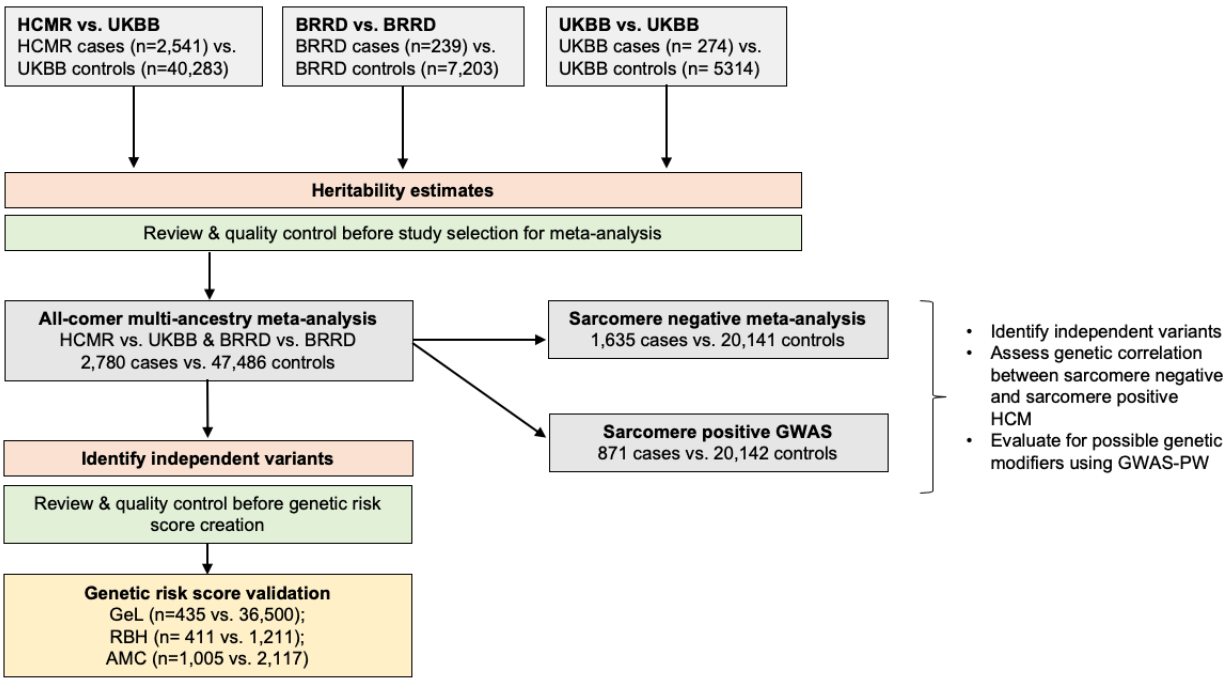


Figure 6.1: Flow diagram outlining GWAS analytical pla BRRD: BioResource for Rare Disease; HCMR: HCM registry; GeL: Genomics England 100,000 Genomes; RBH: Royal Brompton Hospital cohort; AMC: Amsterdam Medical Center

6.2.2 Power calculations

To estimate the probability of rejecting the null hypothesis when it is false, power calculations were performed *a priori* for both the HCMR vs UKBB and the BRRD studies using the University of Michigan Genetic Association Power Calculator (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/).[288]

HCM is considered to represent a dominant trait, however an additive disease model was examined that models a uniform linear increase in risk for every additional common risk allele present within an individual. This is justified as additive models provide closely comparable power to detect both additive and dominant effects and have proven successful in mapping common susceptibility loci for a wide variety of complex traits with unknown penetrance patterns.[289]

The p-value significance threshold used when performing power calculations was 5×10^{-8}. There is consensus that a p-value threshold of 5×10^{-8} minimises false-positive associations without sacrificing true-positive associations, as suggested

by analyses performed by the International HapMap 2 Consortium.[290, 291]

Information regarding risk allele frequencies and genotype relative risk were inferred from the only published HCM GWAS.[118] Although no variants reached genome-wide significance, rs516514, an intronic *FHOD3* variant demonstrates suggestive evidence (OR = 2.45 [95% CI 1.76-3.41], p-value = 1.25×10^{-7}).[118] The global MAF for rs516514 is 0.50 (minimum MAF of 0.42 in the Latino population and a maximum MAF of 0.64 in East Asian ancestries). As HCM is considered a rare disease, genotype relative risk approximates to the OR.

Given the limited size of the BRRD cohort, it was anticipated that the BRRD GWAS would yield modest discovery power: 80% power to correctly reject the null hypothesis for GWAS-significant common variants (i.e. MAF $\geq$0.20) of large effect (i.e. genotype RR $\geq$2), similar to Wooten et al. (2013) (Figure 6.2A). This contrasts with the HCMR vs UKBB GWAS which is 80% powered to detect a GWAS-significant signal, conferring a genotype RR $\geq$2, for variants with a lower allele frequency (i.e. MAF $\geq$0.01) (Figure 6.2B).

6.2.3 Component Genome-wide association studies

HCMR vs UKBB

As outlined in Chapter 2, 2,762 incident HCM cases were recruited from 44 sites across 6 countries in North America and Europe through the HCMR. 2,541 individuals from the HCMR cohort were eligible for analysis, following removal of closely related individuals, (<3rd degree relatives) (n=96), and those with genetically discordant gender from self-reported gender (n=17). 40,283 individuals were quasi-randomly selected from a cohort of 270,260 individuals screened for evidence of cardiomyopathy (described in Chapter 2) for subsequent analysis using a 20:1 allocation against HCMR cases (n=2,541), with approximate age and gender matching.

The HCMR and UKBB cohorts were genotyped on different arrays, specifically the Axiom™Precision Medicine Research Array (Affymetrix/ThermoFisher) and UK Biobank Axiom®array (Affymetrix), respectively. A total of 174,974 genotyped

SNPs (MAF >0.01, genotype missing rate 1%, HWE with mid-p correction of 1×10^{-9}) were present in both the HCMR and UKBB cohorts and taken forward for imputation and further analysis.[160]

To facilitate downstream analysis, the UKBB and HCMR SNPs were aligned to the HRC reference panel, using HRC-1000G-check-bim.pl from https://www.well.ox.ac.uk/~wrayner/tools/ before being merged.

The Michigan Imputation Server (https://imputationserver.sph.umich.edu/) performed haplotype phasing with Eagle, and imputation against the Haplotype Reference Consortium (HRC.r1.1.2016 reference panel), generating genotypes for 38,954,302 imputed variants.[152, 292, 293] Imputed variants with an INFO score >0.3 and MAF >0.01 were retained for subsequent analysis.

PCA was performed using FlashPCA2 on 174,974 genotyped SNPs in approximate linkage equilibrium ($r^2 < 0.05$), determined using the –indep-pairwise function in PLINK (version 1.90b3). Ancestry was inferred by projecting principal components, derived from the 1000 Genomes Project (Phase 3), onto HCMR/UKBB genotypes. A multinomial logistic regression model, performed using the *nnet* CRAN package in R (https://CRAN.R-project.org/package=nnet), classified ancestral groups as specified by the International Genome Sample Resource (http://www.internationalgenome.org/category/population/) (Table 6.1).

All comer analysis was performed using logistic regression to fit an additive case-control association model (cases=2541, controls=40,283), in SNPTEST v2.5.4-beta3, adjusting for the first ten ancestrally informative principal components. As HCM is a disease of a relatively low population prevalence, (∼1 in 500), statistical power was maximised by not adjusting for age or sex.[294]

BRRD vs BRRD

As outlined in chapter 2, 239 sarcomere-negative HCM cases and 7,203 reference controls, enrolled in the BRRD, undertook genome sequencing and were available for analysis. Variants demonstrating: PASS filter status; MAF >1%, a depth of at least 10 informative reads per site; a genotype quality score of at least 20 (GQ >20);

and a genotype missingness of no more than 10% (CR >0.9) were extracted from the BRRD VCF file. Ancestrally informative principal components were derived using FastPCA2 and 1000 Genomes Phase 3 data (Table 6.1). Association analysis was performed using SAIGE (v0.29.4.2) with the first three principal components included as covariates.[295] SAIGE was selected so as to accommodate the presence of related individuals from the BRRD cohort.[296] SAIGE step 1 was performed using 123,903 genotypes following a linkage disequilibrium pruning procedure in PLINK (v1.9), with a 500kb window, a step size of 50 markers, and a pairwise r^2 threshold of 0.2, as described by Zhou et al (2018).[161, 297] SAIGE step 2 analysis was performed using genotypes with a minor allele count >5 and MAF >0.01. Summary genetic association statistics for 9,341,129 autosomal variants were then computed using a mixed-effects logistic regression model with the first three ancestrally informative principal components included as covariates; a genomic control analysis showed little evidence of over-dispersion ($\lambda = 1.049$).

UKBB vs UKBB

Access to the UKBB genotypes was provided through UKBB application 11223. 326 individuals with an ICD10 code indicating HCM (I420 or I421) in HES and available genotyping data performed using the UKBB Axiom Array were considered.

From a possible 229,977 eligible individuals, a random sample of 5,767 individuals were selected, using a 20:1 allocation against UKBB cases, with approximate age (per decade) and gender matching. There was no overlap in samples between this study and the HCMR vs. UKBB study. Principal components were extracted from those centrally generated by the UKBB, with the first ten used as covariates.

The UKBB performed phasing of genotypes using SHAPEIT3 and subsequent imputation was performed using a merged UK10K/1000 Genomes Phase 3 reference panel and the HRC reference panel. Details regarding the genotyping, phasing and imputation are reported elsewhere.[123] SAIGE was used to perform association analysis, with step 1 calibrated using 244,527 genotyped SNPs and step 2 generated

summary statistics for 10,040,434 autosomal SNPs with a with a minor allele count > 5 and minor allele frequency > 0.01.

N (%)	HCMR vs. UKBB		BRRD vs. BRRD		UKBB vs. UKBB		Total	
Ancestry	HCMR cases (n=2,541)	UKBB controls (n=40,283)	Cases (n=239)	Controls (n=7203)	Cases (n=326)	Controls (n=5766)	Cases	Controls
AFR	185 (7.28)	1,032 (2.56)	6 (2.51)	216 (3.00)	21 (6.44)	124 (2.15)	212 (6.83)	1,372 (2.58)
AMR	59 (2.32)	209 (0.52)	2 (0.837)	343 (4.76)	2 (0.613)	20 (0.347)	63 (2.03)	572 (1.07)
EAS	68 (2.68)	310 (0.77)	5 (2.09)	110 (1.53)	1 (0.307)	34 (0.590)	74 (2.38)	454 (0.85)
EUR	2,042 (80.36)	37,274 (92.53)	202 (84.5)	5,394 (74.9)	293 (89.9)	5,441 (94.4)	2,537 (81.7)	48,109 (90.34)
FIN	52 (2.05)	217 (0.54)	0 (0)	81 (1.12)	2 (0.613)	16 (0.277)	54 (1.74)	314 (0.59)
SAS	135 (5.31)	1,241 (3.08)	6 (2.51)	543 (7.53)	7 (2.15)	131 (2.27)	148 (4.76)	1,915 (3.6)
Other	-	-	18 (7.53)	516 (7.16)	-	-	18 (0.58)	516 (0.97)
							3,106	53,252

Table 6.1: Ancestral composition of cases and controls contributing towards HCM GWAS studies Abbreviations: AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; FIN, Finnish European; SAS, South Asian

6.2.4 Heritability estimates

Broad-sense heritability (H^2) refers to the proportion of variation of a trait in a population that can be explained by inherited genetic variants, as denoted through equation I , where σ^2_G is genetic variance and σ^2_P is phenotypic variance. Phenotypic variance and genetic variance (σ^2_G) can be further defined, as shown in equations II and III, respectively. In equation II, σ^2_E denotes environmental variance and $\sigma^2_{E\times G}$ is the variance attributable to gene-environment interactions. The genetic variance, defined in equation III, considers genetic variance from addditive (σ^2_A), dominance (i.e. dominant or recessive effects) (σ^2_D), and epistatic, gene-gene interactions ($\sigma^2_{G\times G}$).

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P} \tag{I}$$

$$\sigma^2_P = \sigma^2_G + \sigma^2_E + \sigma^2_{E\times G} \tag{II}$$

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{G\times G} \tag{III}$$

For common complex diseases, it is widely assumed that the greatest proportion of variance can be attributable to additive effects.[298] Narrow-sense heritability (h^2) intends to estimate the proportion of variation of a trait in a population that can be explained by additive genetic effects, as shown in equation IV. In HCM, where dominant effects are known to exist, h^2 is less than H^2. Traditionally it was only possible to approximate h^2 through closely related individuals (i.e. twin studies), but with advances in GWAS, a proportion of h^2 can be measured in unrelated individuals through commonly detected SNPs (σ^2_{SNPs}). SNP-heritability ($h^2{}_{\mathrm{g}}$) is defined in equation V, where $h^2{}_{\mathrm{g}} \leq h^2 \leq H^2$.

$$h^2 = \frac{\sigma^2_A}{\sigma^2_P} \tag{IV}$$

$$h^2_g = \frac{\sigma^2_{SNPs}}{\sigma^2_P} \tag{V}$$

Here, SNP-heritability ($h^2{}_{\mathrm{g}}$) was approximated using GREML (Genomic relatedness matrix Restricted Maximum Likelihood), part of the Genome-wide Complex Trait Analysis (GCTA) program, for the multi-ancestry HCMR vs UKBB genotypes.[299, 300] The GREML-LDMS method was used to generate a genetic relationship matrix for SNPs demonstrating an INFO score >0.3 and MAF >0.01 (Table 6.2). First LD scores were assigned to SNPs in 200kb regions across the genome, before SNPs were stratified into quartiles based on SNP LD scores and used to generate genomic relationship matrices (GRMs). For each variant in a defined region, the LD score represents the aggregate sum of r^2 values for that specific variant and all other variants in the region. GRMs provide an estimate for the genetic relationships between individuals, based the proportion of alleles shared across a finite number of variants included in the model. Using a restricted maximum likelihood approach, the GRMs are used, alongside the first ten ancestrally informative principal components as covariates, to estimate $h^2{}_{\mathrm{g}}$. $h^2{}_{\mathrm{g}}$ estimates were approximated on a multifactorial liability scale; this assumes that binary traits (i.e. cases vs controls) can be modelled on a continuous scale, and that above a given threshold an individual will be affected. Representation of a binary

trait on the liability scale is dependent on both the population prevalence of disease (0.02% based on population-based epidemiological estimates) and the sample prevalence of disease (i.e. 2,541 cases and 40,283 controls). Of the 0.02%, 40% were assumed to be sarcomere positive and 60% sarcomere negative based on previously published estimates.

BRRD demonstrates the highest ${h^2}_{\mathrm{g}}$ (Table 6.2), and this may reflect a recruitment strategy that prioritised the inclusion of individuals with a family history of HCM and early onset, severe disease, but lacked a pathogenic or likely pathogenic variant across core sarcomere genes. This could conceivably enrich for polygenic contributions towards HCM. Whilst the BRRD ${h^2}_{\mathrm{g}}$ estimates are similar to what has been previously reported for height ($0.602 \pm 4.7 \times 10^{-3}$), there is less precision associated with the BRRD ${h^2}_{\mathrm{g}}$ estimate given the comparatively small sample size (Figure 6.3).[301] The UKBB ${h^2}_{\mathrm{g}}$ estimates appear considerably lower than both those reported for the HCMR and BRRD cohorts. When compared with a range of complex traits, extracted from analysis performed by Hou et al. (2019) in the UKBB, the HCMR cohort appears to confer a relatively high ${h^2}_{\mathrm{g}}$ estimate.[301] This suggests that a larger proportion of phenotypic variability is accounted for by the additive effects of common genetic variants than is observed for other widely studied cardiovascular diseases, such as coronary artery disease. However, it is acknowledged that the missing heritability associated with diseases, such as coronary artery disease, has been established through twin studies, but remains unknown in HCM.[302]

Source	Proportion of cases	Disease prevalence	SNP heritability
UKBB (n=274 cases, 5314 controls)	0.049	0.002	0.026 (0.133)
BRRD (n=211 cases, 4991 controls)	0.0406	0.0012	0.676 (0.164)
HCMR full cohort (n=2541 cases, 40283 controls)	0.0593	0.002	0.352 (0.0181)

Table 6.2: HCM heritability estimates Comparison of SNP heritability estimates derived from multi-ancestry GWAS case-control analyses performed for HCM.

6.3 Methodological considerations for association testing

6.3.1 Genomic control

Genomic inflation is the systemic overdispersion of the test statistic and was assessed through calculation of the genomic control, λ, and by evaluating the overall p-value distribution. Genomic inflation can be attributed to genetic confounding, either through population structure or cryptic relatedness, or polygenicity (i.e. the presence of many true small associations with a phenotype). [156, 303] The genomic control is conventionally used to assess for genomic inflation by measuring the ratio between empirically derived and expected median χ^2 statistics. Genomic control correction procedures can be implemented, typically when $\lambda > 1.1$, to globally reduce false positive associations by down-weighting the reported standard errors (SE) (SE_{gc}= SE $\times \sqrt{\lambda}$) and re-calculating adjusted χ^2 statistics ($\chi^2_{gc} = \left(\frac{\beta}{SE_{gc}}\right)^2$) and associated p-values (under 1 degree of freedom), as proposed by Devlin.[304] However, it is important to appreciate that the genomic control correction cannot differentiate sources of genomic inflation, and when implemented may introduce false negative results. Whilst early GWAS studies typically observed λ values < 1.1, as GWAS studies have incorporated large sample sizes, a λ of >1.1 is increasingly encountered, and is presumably attributable to polygenicity.[305] As an example, the Genetic Investigation of ANthropometric Traits (GIANT) consortium report λ values of 3.6 and 2.7 for height and weight, respectively.[306] To assist with the evaluation of this extreme overdispersion alternative statistical methodologies were developed, such as LD score regression, to help partition true polygenicity from confounding biases.[307]

In this analysis a genomic control correction was performed when λ >1.1.

6.3.2 False discovery rate

GWAS tests for associations between a genetic variant and a trait, across multiple markers simultaneously. The null hypothesis, tested m times, is that each genetic variant is not associated with a trait (i.e. H_0: β=0), as summarised below.

	Non-significant	Significant	Total
H_0	N_{00}	N_{10}	m_0
H_1	N_{01}	N_{11}	m_1
Total	U	R	m

N_{10} represents the total number of type 1 errors (i.e. false positives) and N_{01} represents the number of type 2 errors (i.e. false negatives). The genome-wide significance threshold p-value of 5×10^{-8} was determined using a Bonferroni method that controlled the family-wise error rate (FWER) at an α level of 0.05.[290, 291] The FWER represents the probability of making at least one type 1 error (i.e. $P(N_{10} >0)$. Assuming 1,000,000 tests (m) were performed, to reflect the presumed number of independent (i.e. uncorrelated) common variants across the genome, the FWER would be controlled at an α level of 0.05 with a p-value threshold of 5×10^{-8}.

The introduction of sequencing-based imputation panels markedly increased the number of variants simultaneously assessed using GWAS.[152] However, despite studies suggesting a revised significance threshold (i.e. p-values of 1×10^{-8} for common (MAF $>$ 0.01) variants and 5×10^{-9} for rare (MAF $\leq$0.01) variants imputed from a genome sequencing panel) these recommendations have not yet been widely adopted.[308, 309] It is, however, acknowledged that the Bonferroni procedure generates an overly conservative threshold, given that fewer than 1,000,000 common variants across the genome are likely to be truly uncorrelated due to LD patterns.[310] Consequently, at the expense of aiming to reduce the number of false positive results, the number of false negatives findings may be inflated, particularly when there is low experimental power.

To help address this issue, the false discovery rate (FDR), widely used in other areas of biomedical science, has been applied in the setting of GWAS.[311, 312] The FDR, as developed by Benjamini and Hochberg, denotes the expected proportion of falsely-rejected null hypotheses, or the expected false discovery proportion (i.e. $\frac{N_{10}}{R}$). The FDR is defined as

$$E(N_{10}/R|R > 0)$$

and provides an estimate for the number of correctly-rejected null hypotheses by correcting for anticipated false positives. The FDR has been further developed, through "q-value" and "local FDR" approaches. The q-value is analogous to a p-value, but is derived from the FDR, and set to be the minimum FDR at which the test is called significant by evaluating the full distribution of generated p-values.[313, 314] The local FDR approach is designed to address whether a specific hypothesis, out of the m tests conducted, is true given its p-value.[315] Q values and the local FDR was calculated using the qvalue R package (https://github.com/StoreyLab/qvalue).

6.3.3 Conditional analysis

To identify genetic variants that confer independent risk effects, conditional analysis was performed using a stepwise model selection procedure (–cojo-slct) using GCTA to approximate a multiple logistic regression analysis with more than one associated variant.[299, 316] Summary statistics were supplied from each GWAS study and LD estimates generated from 60,000 unrelated, European individuals randomly selected from the UKBB.

6.3.4 Association analysis results

HCMR vs UKBB

Following genomic control correction, (original $\lambda = 1.191$; summary statistics underwent genomic control adjustment), the multi-ancestry GWAS demonstrated 11 independent genome-wide significant variants, and a further 12 independent variants below the 5% local false-discovery (LFDR) threshold, equivalent to a p-value of 1.5×10^{-6} (Figure 6.4 and Table 6.3).

BRRD vs BRRD

Three SNPs (rs1232572641, rs142939703 and rs16968220) demonstrated genome-wide significance and two (rs61869036 and rs139472654) were below the 5% FDR threshold (p-value=1×10^{-6}) (Figure 6.5, Table 6.4). rs1232572641 and rs142939703 are located 155bp apart, but there is limited evidence to support linkage disequilibrium (r^2= 0.054 and D′=0.362). There was no evidence of overdispersion ($\lambda = 1.049$).

Locus	Chr	SNP	EA	EAF	Summary statistics		
					Beta	SE	*P*-value
BAG3	10	rs72840788	G	0.791	-0.398	0.04	8.14E-24
HSPB7	1	rs1048302	T	0.326	0.275	0.035	1.68E-15
FHOD3	18	rs2644262	T	0.715	-0.298	0.038	1.86E-15
SMARCB1	22	rs7284877	G	0.219	0.293	0.04	1.87E-13
PLN	6	rs12212795	G	0.947	-0.436	0.065	1.98E-11
CDKN1A	6	rs762624	A	0.718	-0.228	0.036	1.94E-10
ADPRHL1	13	rs41306688	A	0.966	-0.589	0.094	3.42E-10
SPPL2C	17	rs393838	G	0.774	-0.242	0.039	4.10E-10
TBX3	12	rs7300371	T	0.267	-0.24	0.041	3.50E-09
PRKCA	17	rs7210446	G	0.419	-0.214	0.038	2.53E-08
OR5AK2	11	rs78310129	C	0.988	-0.775	0.141	3.87E-08
FHOD3	18	rs118060942	C	0.988	-0.645	0.12	7.97E-08
SLC6A6	3	rs13061705	C	0.685	0.225	0.042	9.57E-08
ALPK3	15	rs8033459	C	0.529	-0.181	0.034	1.16E-07
STRN	2	rs11124555	G	0.514	-0.182	0.034	1.19E-07
AK098570	5	rs66761011	A	0.83	-0.349	0.066	1.45E-07
POMT1	9	rs734638	C	0.711	-0.196	0.037	1.54E-07
TRDN	6	rs9320939	G	0.516	-0.176	0.034	2.50E-07
ADAMTS7	15	rs8043123	C	0.757	-0.192	0.038	3.76E-07
SSRP1	11	rs117534260	A	0.981	-0.54	0.108	6.27E-07
FNDC3B	3	rs4894803	A	0.596	-0.186	0.038	8.38E-07
TCF7L2	10	rs11196085	T	0.722	-0.183	0.037	9.39E-07
CYP2R1	11	rs1390519	A	0.666	-0.203	0.042	1.15E-06

Table 6.3: Independent HCMR vs UKBB GWAS results Independent variants as determined by a stepwise model selection procedure using GCTA-COJO. Genome-wide significant loci highlighted in green and 5% LFDR loci highlighted in yellow. Chr = chromosome; SNP = single nucleotide polymorphism; EA = effect allele; EAF = effect allele frequency; SE = standard error; LD = linkage disequilibrium.

UKBB vs UKBB

No genome wide significant loci were detected, but rs74174399 (effect allele = T) (OR: 2.83 [95% CI: 2.42-3.24]; p-value=5.52×10^{-7}) demonstrated association at the 5% FDR threshold (p-value=1×10^{-6}) (Figure 6.6).

6.3.5 Meta-analysis study selection

There is heterogeneity in the HCM GWAS studies, and before proceeding with meta-analysis, each constituent GWAS study was reviewed for suitability. First, in

Locus	Chr	SNP	EA	EAF	Summary statistics		
					beta	se	p-value
.	18	rs1232572641	A	0.05677	1.575	0.234	1.67E-11
.	18	rs142939703	A	0.0294	2.079	0.3274	2.22E-10
FHOD3	18	rs16968220	C	0.6467	-0.5533	0.09966	2.83E-08
BAG3	10	rs61869036	G	0.799	-0.6398	0.1217	1.46E-07
YTHDC2	5	rs139472654	C	0.97662	-1.806	0.349	2.28E-07

Table 6.4: Independent BRRD vs BRRD GWAS results Independent variants as determined by a stepwise model selection procedure using GCTA-COJO. Genome-wide significant loci highlighted in green and 5% LFDR loci highlighted in yellow. Chr = chromosome; SNP = single nucleotide polymorphism; EA = effect allele; EAF = effect allele frequency; SE = standard error; LD = linkage disequilibrium.

terms of recruitment and case definitions, the HCMR and BRRD studies appear superior to the UKBB study. Individuals recruited to both HCMR or BRRD cohorts were clinically evaluated by a cardiologist with expertise in ICCs. This contrasts with the UKBB, which included individuals based on relatively sparse information derived from ICD10 codes, which may not be accurate at present, and therefore, renders them unsuitable (see Chapter 5).

Supporting the decision to exclude the UKBB GWAS from the meta-analysis based on uncertain inclusion criteria is the observation of limited h^2_g, which appears to be incongruent to the h^2_g derived from the BRRD and HCMR cohorts.

6.3.6 Meta-analysis

A fixed-effects inverse-variance meta-analysis was implemented using GWAMA for two multi-ancestry genome-wide association studies (HCMR cases (n=2,541) vs. UKBB (n=40,283); and BRRD cases (n=239) vs. BRRD controls (n=7,203)) using 8,590,397 SNPs.[317]

A fixed-effects model assumes the true effect size is the same across all included studies, and differs from a random-effects model which allows the true effect to vary between studies. The inverse-variance meta-analysis method is performed for each variant, with allelic effects weighted by the inverse of their variance, as derived from the 95% confidence interval of the OR, across each study.[317]

Locus number	Chr	SNP	Position (GRCh37)	NEA/ EA	Freq EA	OR [95% CI]	P-value	Locus name
GWAS significant ($P < 5 \times 10^{-8}$)								
1	1	rs1048302	16340879	G/T	0.33	1.32 [1.24-1.40]	2.54E-17	*HSPB7*
2	3	rs13061705	14291129	T/C	0.69	1.25 [1.16-1.34]	9.18E-09	*SLC6A6*
3	6	rs3176326	36647289	G/A	0.21	1.28 [1.19-1.38]	2.22E-11	*CDKN1A*
4	6	rs12212795	118654308	G/C	0.05	1.48 [1.31-1.67]	2.51E-10	*PLN*
5	10	rs72840788	121415685	G/A	0.21	1.52 [1.42-1.64]	5.06E-29	*BAG3*
6	12	rs7301677	115381147	T/C	0.73	1.24 [1.15-1.33]	1.26E-08	*TBX3*
7	13	rs41306688	114078558	A/C	0.03	1.82 [1.53-2.17]	1.08E-11	*ADPRHL1*
8	15	rs8033459	85253258	C/T	0.47	1.21 [1.14-1.29]	3.41E-09	*ALPK3*
9	17	rs28768976	43688317	A/G	0.23	1.29 [1.20-1.39]	4.12E-12	*SPPL2C*
10	17	rs7210446	64307014	G/A	0.58	1.25 [1.16-1.34]	6.82E-10	*PRKCA*
11	18	rs4799426	34280891	A/G	0.35	1.38 [1.29-1.47]	4.00E-23	*FHOD3*
11	18	rs118060942	34280732	C/T	0.01	2.22 [1.78-2.78]	3.23E-12	*FHOD3*
12	22	rs2070458	24159307	T/A	0.22	1.34 [1.25-1.44]	7.12E-15	*MMP11*
LFDR < 5% ($P < 1.82 \times 10^{-6}$)								
13	2	rs7556984	11599732	A/G	0.66	1.20 [1.12-1.29]	5.21E-7	*E2F6, ROCK2*
14	2	rs2003585	37123383	C/T	0.49	1.19 [1.12-1.27]	8.60E-8	*STRN*
15	2	rs62177303	179768624	T/C	0.61	1.19 [1.11-1.28]	7.00E-7	*TTN*
16	3	rs4894803	171800256	A/G	0.40	1.20 [1.12-1.28]	3.51E-7	*FNDC3B*
17	5	rs66761011	29426502	A/G	0.17	1.42 [1.24-1.61]	1.45E-7	*AK098570*
18	5	rs10052399	138668504	C/T	0.26	1.23 [1.14-1.32]	6.21E-8	*PROB1*
19	6	rs9320939	123818871	G/A	0.48	1.19 [1.12-1.27]	5.78E-08	*TRDN*
20	8	rs7003871	125851510	T/C	0.34	1.18 [1.11-1.27]	1.14E-6	*MTSS1*
21	9	rs734638	134489810	C/G	0.29	1.20 [1.12-1.29]	1.09E-7	*RAPGEF1, POMT1*
22	10	rs11196085	114505037	T/C	0.28	1.21 [1.13-1.30]	7.30E-8	*TCF7L2*
23	11	rs1390519	14928023	A/G	0.33	1.22 [1.13-1.33]	1.16E-6	*CYP2R1*
24	11	rs78310129	56793878	C/T	0.012	2.16 [1.64-2.84]	5.17E-08	*OR5AK2*
25	12	rs1480036	26344726	T/C	0.76	1.23 [1.14-1.34]	3.50E-7	*SSPN*
26	15	rs1814880	79021140	C/T	0.26	1.20 [1.12-1.28]	1.59E-7	*CHRNB4*
27	19	rs117710064	779606	C/T	0.14	1.25 [1.14-1.36]	8.55E-7	*AZU1*
28	21	rs2832230	30536712	T/G	0.83	1.29 [1.16-1.42]	9.57E-7	*MAP3K7CL*

Table 6.5: Meta-analysis results: variants independently associated with HCM beneath the genome-wide significant and 5% local false discovery rate threshold. Abbreviations: CI: confidence interval; EA: effect allele; FDR: false discovery rate; Freq EA: effect allele frequency; GWAS: genome-wide association study; NEA: Non-effect allele; OR: odds ratio; SNP: single nucleotide polymorphism.

The HCMR vs UKBB underwent a genomic control procedure prior to being included in the fixed effects inverse-variance meta-analysis. There was no evidence of over dispersion ($\lambda = 1.03$). 28 discrete loci were identified, containing 13 independent genome-wide significant variants and 16 independent LFDR variants (p-value threshold = 1.82×10^{-6} and q-value threshold = 4.18×10^{-3}) (Figure 6.7, Table 6.5). The *FHOD3* locus was found to harbour two independent genome-wide significant variants, rs4799426 and rs118060942, in linkage equilibrium (r^2=0.01).

6.3.7 Locus assessment

Most risk-associated variants that have been detected via genome-wide association analysis are located in non-coding genomic regions, particularly cell-type-specific regulatory elements.[318] Consequently, the process of assigning causality to candidate genes remains a challenging process. Several methods have been posited, including: proximity to a neighbouring transcription start site; colocalization of association signals between a trait/disease and molecular quantitative trait loci; evidence of monogenic disease in a neighbouring gene; missense variants in LD with risk-associated SNP; identification of tissue-specific molecular quantitative trait loci (QTLs) and the identification of tissue-specific chromosomal conformation.[319, 320]

Risk-associated loci derived from the multi-ancestry meta-analysis were evaluated in an attempt to identify genes influencing susceptibility towards HCM using several complementary approaches. dbSNP was consulted to identify the variant class and allele frequency in external population databases. To assess expression QTLs (eQTLs) and physical chromosomal interactions, in myocardial tissue, summary statistics were assessed in FUMA v1.3.5e (http://fuma.ctglab.nl) using tissue expression data from GTEx v8.0 (https://www.gtexportal.org/home) and chromatin interaction Hi-C data from Schmitt et al (2016).[321, 322] OpenTargets (https://genetics.opentargets.org/) was used to identify proximity to transcription start sites, protein QTL (pQTL) associations and associations with additional phenotypes. Several of the HCM risk-associated variants demonstrated supporting evidence to suggest a plausible candidate gene, including: *BAG3, FHOD3, HSPB7, ADPRHL1* and *SLC6A6*.

BAG3

An intronic variant, rs72840788 (risk allele = A; effect allele frequency = 0.21; OR= 1.52 [95% CI: 1.41 – 1.63]; p-value = 5.06×10^{-29}), in the *BAG3* (BLC2-associated athanogene 3) locus, demonstrated association with HCM. When conditional analysis was performed, adjusting for the genetic effects of the sentinel SNP (rs72840788), no additional independent SNPs were detected.

BAG3 is widely expressed, including in muscle and heart tissue. Rare missense variants in *BAG3*, specifically p.Pro209Leu, have been detected in individuals diagnosed with a myofibrillar myopathy who had evidence of either hypertrophic or restrictive cardiomyopathy.[323, 324] rs72840788 is in LD with a *BAG3* missense variant, rs2234962 ($D' = 1.00$ and $r^2 = 0.99$), that demonstrates a phred-scaled Combined Annotation Dependent Depletion (CADD) score of 21.5, suggesting a deleterious impact. Furthermore, whilst supportive cis-eQTL or promoter capture Hi-C data implicating rs72840788 with *BAG3* is lacking, pQTL data indicates that rs72840788 increases BAG3 protein levels in blood plasma (β: 0.177; p-value= 5.1×10^{-9}).[325, 326]

BAG3 is a member of the BAG family of anti-apoptotic proteins (BAG 1-6), that bind and regulate the activity of heat shock protein 70 (Hsp70) via a BAG domain in the C-terminal. Hsp70 is a ubiquitously expressed protein that regulates the quality of protein folding. Given this, it is intriguing that rs1048302, a 3′ UTR variant in *HSPB7*, is significantly associated with HCM (OR 1.32 [95% CI: 1.24-1.40]; p-value = 2.51×10^{-17}). Within the cardiomyocyte, *BAG3* contributes towards the maintenance of sarcomeric structural integrity during mechanical stress, acts as a homeostatic regulator of filamin, and contributes towards the removal of misfolded or degraded protein products through chaperone-assisted selective autophagy (CASA).[327]

Phenome-wide association analysis for rs72840788, or SNPs in LD with rs72840788, associate with myocardial biology related traits and diseases, including: heart failure (rs17617337; in LD with rs72840788 ($D' = 1.00$ and $r^2 = 0.99$)), dilated cardiomyopathy (rs2234962; in LD with rs72840788 ($D' = 1.00$ and $r^2 = 0.99$)), reduced left-ventricular end systolic volume (rs72840788) and increased left-ventricular ejection fraction (rs72840788).[200, 227, 328]

FHOD3

An intronic variant, rs4799426 (risk allele = G; effect allele frequency = 0.35; OR= 1.35 [95% CI: 1.26 – 1.43]; p-value = 4×10^{-23}), in the *FHOD3* (formin

homology 2 domain containing 3) locus, demonstrated association with HCM. Conditional analysis, that accounted for the genetic effects of the sentinel SNP (rs4799426), revealed rs118060942 as an additional, independent SNP (risk allele = T; effect allele frequency = 0.012; OR = 1.95 [95% CI: 1.51-2.29]; p-value = 7.07×10^{-9}) within the *FHOD3* locus.

FHOD3 appears to be a strong candidate gene for HCM associated risk, having demonstrated evidence of co-segregation for numerous pathogenic, rare variants, and association with HCM in a previous GWAS.[65, 118] Additionally, rs879568, a *FHOD3* intronic variant in LD with the sentinel SNP (rs4799426, $r^2 = 0.90$) has demonstrated an association with QRS-duration (p-value= 8.0×10^{-9}).[329]

Additionally, rs4799426 is in LD ($r^2 = 0.84$) with rs2303510, a *FHOD3* missense variant, with a CADD score of 22.8. Whilst rs4799426 does not appear to directly influence *FHOD3* expression, SNPs in LD with rs4799426 (rs1495900, $r^2 = 1.0$, GWAS p-value = 2.84×10^{-22}) do appear to influence *FHOD3* in heart atrial appendage tissue (normalised effect size = -0.15; p-value = 4.2×10^{-8}). Formal colocalisation analysis is required to further evaluate overlapping signals within the *FHOD3* locus. *FHOD3* appears highly expressed in cardiac tissue and contributes towards the organisation of actin and in the maintenance of cardiac function.[330] Functional studies in mice suggest *FHOD3* directly interacts with *MYBPC3*, a mechanism that appears to be important for the regulation of cardiac function.[331]

HSPB7

rs1048302, a 3′ UTR variant in *HSPB7* (heat shock protein family B (small) member 7) demonstrated association with HCM (risk allele = T; effect allele frequency = 0.33; OR = 1.32 [95% CI: 1.24 – 1.40]; p-value = 2.51×10^{-17}). Conditional analysis, accounting for the genetic effects of the sentinel SNP (rs1048302), revealed no additional independent SNPs. *HSPB7* is highly expressed in cardiac and muscle tissue, and SNPs in LD with rs1048302 appear to influence *HSPB7* expression in atrial appendage tissue.

rs1048302 is in LD with SNPs associated with DCM (rs10927875, $r^2 = 0.80$; p-value = 1.0×10^{-9}) and there is suggestive evidence supporting an association with heart failure.[200, 328, 332]

HSPB7 has a critical role in cardiac development, with *HSPB7* knockout mouse models demonstrating embryonic lethality, with evidence of smaller heart size and congestive heart failure.[308] *HSPB7* appears to have several roles in maintaining muscle integrity, with roles in modulating the actin thin filament length and consequent suppression of actin polymerisation, but also through an interaction with dimerized *FLNC*.[333]

ADPRHL1

rs41306688, a missense variant with a CADD score of 27.1 in *ADPRHL1*, (ADP-ribosylhydrolase like 1) was proven to be associated with HCM (risk allele = C; effect allele frequency = 0.03; OR = 1.82 [95% CI: 1.51 – 2.09]; p-value = 1.06×10^{-11}). In a meta-analysis of 22 studies, rs41306688 has been shown to prolong PR interval (β = 0.10; standard error=0.02; p-value = 7.4×10^{-9}).[334]

ADPRHL1 appears highly expressed in cardiac and muscle tissue. Functional studies, performed in Xenopus, suggest *ADPRHL1* has a critical role in modifying Z-disc and actin dynamics during cardiac development.[335]

SLC6A6

rs13061705 is associated with risk of HCM (risk allele = C; effect allele frequency = 0.69; OR = 1.25 [95% CI: 1.16 – 1.34]; p-value = 9.06×10^{-9}) and is located in an inter-genic region. *Cis*-eQTL analysis from heart atrial appendage tissue in GTEx (v8) indicates that SNPs in LD with rs13061705 influence the expression of *SLC6A6*. For example, rs62231954 (rs13061705, $r^2 = 0.53$) increases *SLC6A6* in heart atrial appendage tissue (normalised effect size = 0.18, p-value = 8.5×10^{-5}). *SLC6A6* encodes a taurine transporter that appears ubiquitously expressed. Numerous animal studies have demonstrated that taurine deficiency leads to a DCM phenotype.[336–339]

6.4 The genetic architecture of sarcomere positive and sarcomere negative HCM

6.4.1 Dichotomise HCM based on sarcomere variant carrier status

Given the considerable differences in h^2_g estimates provided by an all-comer HCMR cohort analysis and the sarcomere negative BRRD cohort, it was hypothesised that sarcomere negative HCM conferred a higher h^2_g estimate than sarcomere positive HCM. If proven true, it could be further hypothesised that the genetic architecture of sarcomere positive and sarcomere negative HCM are distinct from one another.

To formally evaluate these hypotheses, HCM cases from the HCMR cohort were dichotomised into sarcomere-positive and sarcomere-negative groups, with an evidence-based gene-specific framework used to parse VUSs.[8] Comparing groups partitioned by sarcomere status afforded greater statistical power than an analysis split by ACMG/AMP categorical groups, due to the relative increase in samples per group assessed. This approach could be considered to be more clinically meaningful, given that VUSs confer clinical-outcomes that closely resemble pathogenic variants.[56] Furthermore, given the large excess of VUSs in HCM cases, and clinical outcomes data, it could be speculated that two-thirds of VUSs are in fact pathogenic.[8]

Consequently, VUSs (n=338) were parsed into one of three groups: VUS-favours benign (n=48, 14.2%), VUS-indeterminate (n=215, 63.6%) or VUS-favours pathogenic (n=74, 21.9%); using a gene-specific framework (see Chapter 2). The cumulative contribution of each VUS group, in combination with likely pathogenic and pathogenic variants, was directly compared against an independent case-control analysis.[8] This indicated that for *ACTC1, MYH7, MYL2, MYL3, TNNT2, TNNI3* and *TPM1*, the inclusion of VUS-indeterminate and VUS-favours pathogenic categories best approximated the expected frequency of causal mutations (Table 6).[8] For *MYBPC3*, an excess of missense variants in the VUS-indeterminate group were too numerous to be considered causal of HCM. Consequently, individuals were included in the sarcomere positive group if they harboured a variant classified as

		Cumulative Counts (n (%))				Excess of rare variants (ExAC MAF < 1×10^{-4}) between HCM cases and ExAC controls
Gene	Variant group	VUS (FB), VUS (I), VUS (FP), LP, P	VUS (I), VUS (FP), LP, P	VUS (FP), LP, P	LP, P	
MYBPC3	all variants	614 (23.29)	601 (22.80)	497 (18.85)	488 (18.51)	17.06%
MYBPC3	truncating	255 (9.67)	255 (9.67)*	253 (9.60)	252 (9.56)	9.07%
MYBPC3	non-truncating	359 (13.62)	346 (13.13)	244 (9.26)*	236 (8.95)	7.99%
MYH7	all variants	320 (12.14)	308 (11.68)*	240 (9.10)	212 (8.04)	12.82%
TNNI3	all variants	48 (1.82)	47 (1.78)*	41 (1.56)	34 (1.29)	2.01%
TNNT2	all variants	37 (1.40)	31 (1.18)*	21 (0.80)	18 (0.68)	1.71%
MYL2	all variants	35 (1.33)	24 (0.91)*	13 (0.49)	13 (0.49)	0.92%
TPM1	all variants	30 (1.14)	27 (1.02)*	24 (0.91)	7 (0.27)	1.40%
MYL3	all variants	18 (0.68)	16 (0.61)*	4 (0.15)	4 (0.15)	0.70%
ACTC1	all variants	12 (0.46)	12 (0.46)*	11 (0.42)	1 (0.04)	0.46%
Total		1114	1066	851	777	

Table 6.6: Gene-based, evidence driven, approach for the parsing of variants of uncertain significance Cumulative counts for all potentially disease causing variants directly compared with the excess of rare variants observed in an independent case-control analysis to facilitate the stratification of HCM cases by sarcomere status. *Cells highlighted in pink show groups that approximately match the excess observed in Walsh et al. (2016)[8]

either VUS-indeterminate, VUS-favours pathogenic, likely pathogenic or pathogenic in *ACTC1, MYH7, MYL2, MYL3, TNNT2, TNNI3* and *TPM1*, or a VUS-favours pathogenic, likely pathogenic or pathogenic in *MYBPC3*. Overall, the HCMR cases were stratified into either a sarcomere positive (n=943) or sarcomere negative (n=1,693) group. Subsequent analyses stratified by sarcomere status was performed, including phenotypic evaluation, GWAS and heritability estimates.

6.4.2 Phenotypic evaluation of HCM stratified by sarcomere status

Comparative analysis evaluating differences between the sarcomere positive and sarcomere negative HCM cohorts was performed by the central HCMR group.[340] In brief, individuals from the sarcomere negative group demonstrated less gadolinium enhancement, less interstitial fibrosis and more left ventricular obstruction than sarcomere positive HCM (Table 6.7). Further details of the analysis performed can be found in Neubauer et al. (2019).[340]

	Sarcomere positive (n=943, 35.8%)	Sarcomere negative (n=1693, 64.2%)	Unadjusted P-value
Age, yrs	46.2 ± 12.0	51.3 ± 10.4	<0.001
BMI, kg/m^2	28.2 ± 5.4	29.8 ± 5.6	<0.001
Male (%, (n))	65.1 (611)	75.3 (1,260)	<0.001
Minority (%, (n))	28.4 (116)	37.4 (823)	0.001
Family history of HCM (%, (n))	54.7 (511)	22.3 (371)	<0.001
PCA derived Ancestry			
AFR	49 (5.63)	145 (8.87)	
AMR	11 (1.26)	33 (2.02)	
EAS	17 (1.95)	46 (2.81)	
EUR	729 (83.79)	1284 (78.53)	
FIN	22 (2.53)	39 (2.39)	
SAS	42 (4.83)	88 (5.38)	
LVOT ≥ 30 mmHg (%, (n))	19.0 (130)	26.8 (335)	<0.001
Arrhythmias (%, (n))	38.1 (185)	35.4 (749)	0.255
Hypertension (%, (n))	21.3 (199)	45.1 (752)	<0.001
LVEF < 55% (%, (n))	14.2 (126)	14.2 (227)	0.983
Morphology:			
Reverse curvature asymmetric septal hypertrophy	58.1%	30.7%	
Isolated basal septal hypertrophy	33.8%	51.8%	
Apical hypertrophy	4.5%	10.7%	
Concentric hypertrophy	0.2%	2.0%	
Other hypertrophy	0.8%	1.6%	
Mid-cavity obstruction with apical aneurysm	2.6%	3.2%	<0.001
Extracellular volume	0.30 ± 0.05	0.29 ± 0.05	<0.01
Native T1			
1.5 T	978 ± 76	968 ± 74	
3.0 T	1,175 ± 89	1,167 ± 81	0.21
Late gadolinium enhancement (LGE) (%, (n))			
No LGE (n = 1,213)	30.1 (264)	60.3 (949)	
0%–5% (n = 965)	52.2 (458)	32.2 (507)	
5%–10% (n = 177)	11.5 (101)	4.8 (76)	
10%–15% (n = 53)	3.8 (33)	1.3 (20)	
>15% (n = 45)	2.5 (22)	1.5 (23)	

Table 6.7: Phenotypic characteristics of individuals enrolled in the HCMR cohort, stratified by sarcomere status. Abbreviations: BMI: body mass index; LVEF: left ventricular ejection fraction; LVOT: left ventricular outflow tract.

6.4.3 SNP heritability for HCM stratified by sarcomere status

From the HCMR vs UKBB cohort, sarcomere negative individuals (n= 1,635) and sarcomere positive individuals (n=871) were identified. UKBB controls used in the all-comer HCMR vs UKBB GWAS were randomly allocated to either the sarcomere negative (n=20,141) or sarcomere positive (n=20,142) GWAS. SNPTEST was used to perform both the sarcomere negative and sarcomere positive GWAS. A genomic control correction was applied to the sarcomere negative HCMR GWAS (pre-λ_{GC}= 1.14, post-λ_{GC}=1.0), but this was not necessary for the sarcomere positive GWAS

Source	Proportion of cases	Disease prevalence	h^2_g(SE)
HCMR full cohort (n=2541 cases, 40283 controls)	0.0593	0.002	0.352 (0.018)
HCMR sarcomere positive (n=871 cases, 20142 controls)	0.0415	0.0008	0.158 (0.038)
HCMR sarcomere negative (n=1635 cases, 20141 controls)	0.0751	0.0012	0.340 (0.024)

Table 6.8: HCM heritability estimates split by sarcomere variant carrier status Abbreviations: h^2_g, SNP heritability; SE, standard error

($\lambda = 1.09$), as there was no evidence of extreme genomic inflation.

SNP heritability for HCM, stratified by sarcomere status, was estimated using GREML-LDMS. The sarcomere negative cohort appeared to demonstrate higher h^2_g (34.0±2.4%) than the sarcomere positive cohort (15.8±3.8%) (Table 6.8). This observation is consistent with h^2_g estimate findings from the BRRD cohort and suggests common genetic variants are a major contributor towards sarcomere negative HCM. The all-comer HCMR GWAS demonstrated h^2_g estimates of 35.2±1.8%, which is similar to findings from the sarcomere negative HCMR vs UKBB analysis.

6.4.4 Comparing the genetic architectures of sarcomere positive and sarcomere negative HCM

Given the observed differences, in both h^2_g estimates and reported GWAS loci, between sarcomere positive HCM and sarcomere negative HCM, bivariate GREML analysis was performed to assess the correlation between the two traits. As individual level data were available bivariate GREML analysis was pursued, rather than analyses that are contingent on summary-level information, such as LD score regression.[307] Bivariate GREML analysis uses a linear mixed model to approximate both the genetic variances between sarcomere positive and sarcomere negative HCM, and the covariance between them. Overall, there was a strong positive correlation ($r_g = +1\pm0.12$) between sarcomere positive and sarcomere negative HCM (Table 6.9).

	Sarcomere negative	Sarcomere positive	Combined
Genetic variance (V(G))	0.014 (0.001)	0.004 (0.001)	-
Residual variance (V(E))	0.047 (0.001)	0.036 (0.001)	-
Genetic covariance (C(G))	-	-	0.008 (0.001)
Phenotypic variance (Vp)	0.061 (0.001)	0.040 (0.000)	-
h^2_g (V(G)/Vp)	0.233 (0.018)	0.101 (0.019)	-
h^2_g on liability scale (V(G)/Vp_L)	0.304 (0.024)	0.217 (0.040)	-

Genetic correlation (r_g (SE))	1.00 (0.123)
logL	-43294.079
n	42789

Table 6.9: Genetic correlation analysis between sarcomere negative and sarcomere positive HCM Approach deploys bivariate GREML analysis using the HCMR cohort.

6.4.5 HCM GWAS stratified by sarcomere status

For the sarcomere positive HCM vs. UKBB GWAS, a conditional and joint analysis was performed using GCTA-cojo and identified 7 independent SNPs at genome-wide significance alongside 13 independent SNPs beneath the 5% LFDR threshold (p-value = 1.59×10^{-6}) (Figure 6, Table 10). GCTA-cojo is an established software package that performs conditional and joint multiple-SNP analyses, using GWAS summary statistics, for the identification of SNPs that confer independent effects[299]

The sarcomere negative HCMR vs UKBB GWAS was meta-analysed with the BRRD vs BRRD GWAS using GWAMA. Using GCTA-cojo, 12 independent SNPs at genome-wide significance were identified alongside 13 independent SNPs beneath the 5% LFDR threshold (p-value = 1.56×10^{-6}) (Figure 6.8, Table 6.10).

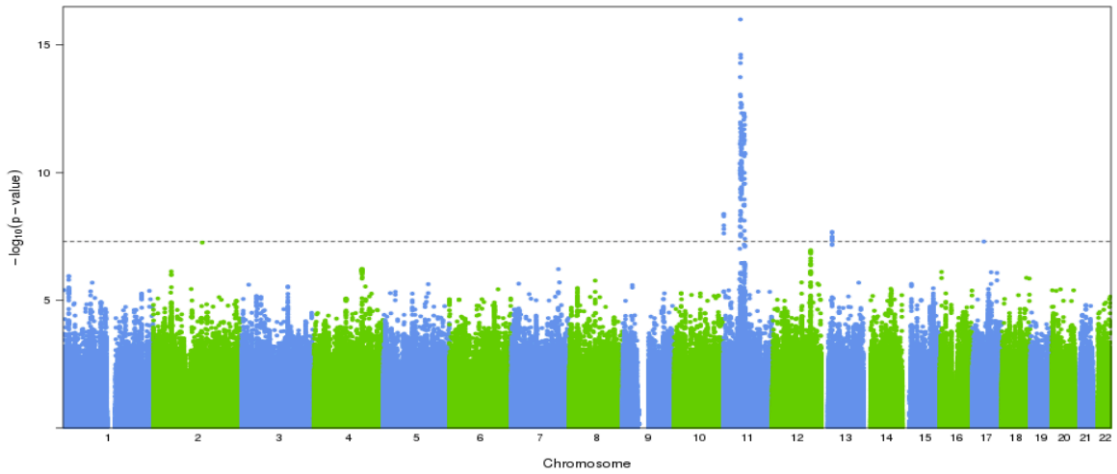


Figure 6.8: Sarcomere positive HCMR vs UKBB GWAS results represented using a Manhattan plot Multi-ancestry Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}).

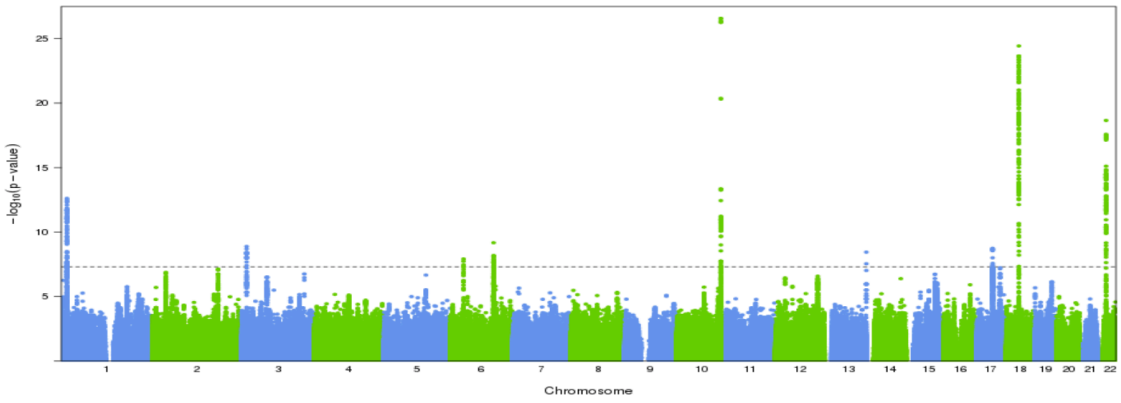


Figure 6.9: Sarcomere negative meta-analysis results represented using a Manhattan plot Multi-ancestry Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}).

6.4.6 Identifying the shared genetic influences across sarcomere positive and sarcomere negative HCM

To complement the bivariate GREML analysis, GWAS-PW was performed by Dr. Anuj Goel (Wellcome Centre for Human Genetics, University of Oxford) using a Bayesian framework outlined by Pickrell et al (2016).[341] GWAS-PW aims to

Chr	SNP	POS	EA	EAF	BETA	SE	p-value
GWAS significant							
11	rs142929136	46513743	T	0.987	-1.306	0.157	9.74E-17
11	rs139913353	56701515	T	0.988	-1.248	0.173	4.66E-13
11	rs74605438	45522618	C	0.987	-1.194	0.174	7.37E-12
11	rs78631951	57917265	C	0.984	-1.012	0.151	1.88E-11
11	rs34225581	46097595	A	0.932	-0.547	0.089	9.23E-10
11	rs35469308	1016869	A	0.84	-0.577	0.098	4.21E-09
13	rs2758215	29537036	G	0.091	0.434	0.077	2.10E-08
<5% LFDR							
2	rs189654928	133777292	C	0.989	-1.014	0.187	5.51E-08
12	rs12299450	104658886	T	0.946	-0.537	0.101	1.10E-07
11	rs117461622	49110043	T	0.987	-0.952	0.187	3.47E-07
4	rs2312403	131847994	C	0.438	-0.282	0.056	5.87E-07
7	rs76600480	128549878	C	0.974	-0.69	0.138	6.00E-07
2	rs35258848	48851517	T	0.871	-0.364	0.073	7.44E-07
16	rs60247077	5287275	T	0.751	-0.275	0.056	7.82E-07
17	rs186910954	51364578	A	0.984	-0.971	0.197	7.96E-07
17	rs76905625	67322589	C	0.985	-0.848	0.172	8.38E-07
1	rs2015509	16358148	T	0.329	0.262	0.054	1.12E-06
11	rs11038225	44976681	G	0.91	-0.433	0.09	1.53E-06
11	rs151252290	57076524	C	0.981	-0.714	0.151	2.17E-06
11	rs116897257	57137809	C	0.977	-0.637	0.137	3.13E-06

Table 6.10: Independent loci associated with sarcomere positive HCM

identify whether regions identified in one GWAS study also contribute towards association signals in another GWAS study (Figure 6.10).

Chr	SNP	POS	EA	EAF	BETA	SE	p-value
GWAS significant							
1	rs1048302	16340879	T	0.326	0.291	0.04	2.36E-13
2	rs17362588	179721046	G	0.916	-0.346	0.064	7.39E-08
3	rs9843704	14297304	C	0.783	0.321	0.053	1.27E-09
6	rs3176326	36647289	G	0.788	-0.255	0.045	1.19E-08
6	rs28436726	118664854	G	0.938	-0.437	0.071	6.50E-10
10	rs72840788	121415685	G	0.792	-0.492	0.045	2.47E-27
13	rs41306688	114078558	A	0.967	-0.63	0.107	3.53E-09
17	rs28768976	43688317	A	0.771	-0.268	0.045	1.77E-09
17	rs7210446	64307014	G	0.418	-0.233	0.043	5.71E-08
18	rs617207	34244842	A	0.291	0.43	0.041	3.41E-25
18	rs118060942	34280732	C	0.988	-1.036	0.132	4.06E-15
22	rs6003909	24181652	A	0.19	0.406	0.045	2.08E-19
<5% LFDR							
1	rs846111	6279370	G	0.733	0.255	0.051	5.50E-07
2	rs2003585	37123383	T	0.493	0.207	0.039	1.33E-07
3	rs62253180	69906572	A	0.817	-0.255	0.05	3.05E-07
3	rs4894803	171800256	A	0.596	-0.222	0.042	1.73E-07
5	rs73780096	114621011	C	0.937	0.448	0.086	2.13E-07
6	rs9320939	123818871	G	0.517	-0.186	0.039	1.54E-06
12	rs17380837	26345526	C	0.711	0.23	0.045	3.52E-07
12	rs7977151	115362972	A	0.304	-0.228	0.044	2.58E-07
12	rs35357	115606207	A	0.171	0.254	0.052	8.46E-07
15	rs8043123	78973393	C	0.754	-0.224	0.043	1.84E-07
15	rs748455	85149575	T	0.716	0.22	0.045	8.75E-07
16	rs139235535	72759681	C	0.975	-0.549	0.113	1.19E-06
19	rs8106955	46346330	T	0.312	-0.21	0.042	7.19E-07

Table 6.11: Independent loci associated with sarcomere negative HCM Results from the meta-analysis of sarcomere negative HCM from HCMR vs UKBB and BRRD vs BRRD

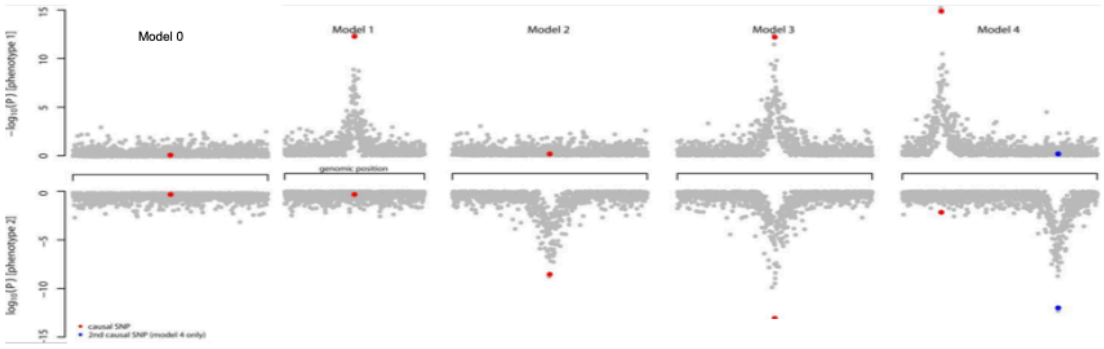


Figure 6.10: Schematic highlighting the different models evaluated by GWAS-PW SNPs demonstrating association for either sarcomere negative or sarcomere positive HCM were directly compared to identify whether the SNP contributed towards: neither trait (model 0), one or the other traits (model 1 or 2), both traits (model 3), or if two independent signals in the same locus contribute towards both traits separately (model 4)

Application of GWAS-PW to the HCM analysis, stratified by sarcomere variant carrier status, suggests that 59% (n=22/37) of regions, identified by the presence of a sarcomere positive or sarcomere negative associated variant ($< 5\%$ LFDR threshold), contributes towards both sarcomere positive and sarcomere negative HCM (Table 6.12). Four SNPs were identified as contributing towards an association signal in the sarcomere positive GWAS, but not in the sarcomere negative GWAS (rs2312403, rs35469308, rs12299450 and rs2758215) and may represent disease modifiers.

6.4.7 Power analysis: sarcomere negative loci in the sarcomere positive GWAS

Given the strong positive genetic correlation between sarcomere positive and sarcomere negative HCM, supported by both bivariate GREML analysis and GWAS-PW, I hypothesised that many of the genome-wide significant variants identified in the sarcomere negative analysis would not yield a comparatively extreme p-value in the sarcomere positive analysis due to limited discovery power.

To evaluate this, power to detect a signal in the sarcomere positive GWAS was calculated using variants identified in the sarcomere negative GWAS (Table 6.13). Of the 26 variants evaluated, only 3 variants identified through the sarcomere negative GWAS (rs72840788, rs617207 and rs6003909) were adequately powered

(i.e. power > 80%) in the sarcomere positive GWAS to detect a signal. This suggests the sarcomere positive GWAS was systemically underpowered to validate the sarcomere negative results.

6.4.8 Assessment for synthetic association with rare, pathogenic variants in MYBPC3

The sarcomere positive HCMR vs UKBB GWAS yields five independent (i.e. in approximate linkage equilibrium) genome-wide significant SNPs within the pericentromeric region of chromosome 11, specifically between positions 40Mb and 60Mb (GRCh37) (Figure 6.11). Variants spanning positions 40Mb to 60Mb on chromosome 11 have previously been shown to associate with HCM in a GWAS performed using 192 cases and 436,274 controls from FINNGEN and the UKBB (Figure 6.12 and Table 6.14).[342] This pericentromeric region of chromosome 11, whilst identified as harbouring long-range, high linkage disequilibrium (a possible consequence of suppressed of meiotic recombination given physical proximity to the centromere), also contains *MYBPC3*.[343–346] Therefore, it is plausible that the numerous "independent" loci detected across chromosome 11 (40-60Mb) are in synthetic association with rare, pathogenic variants in *MYBPC3*.

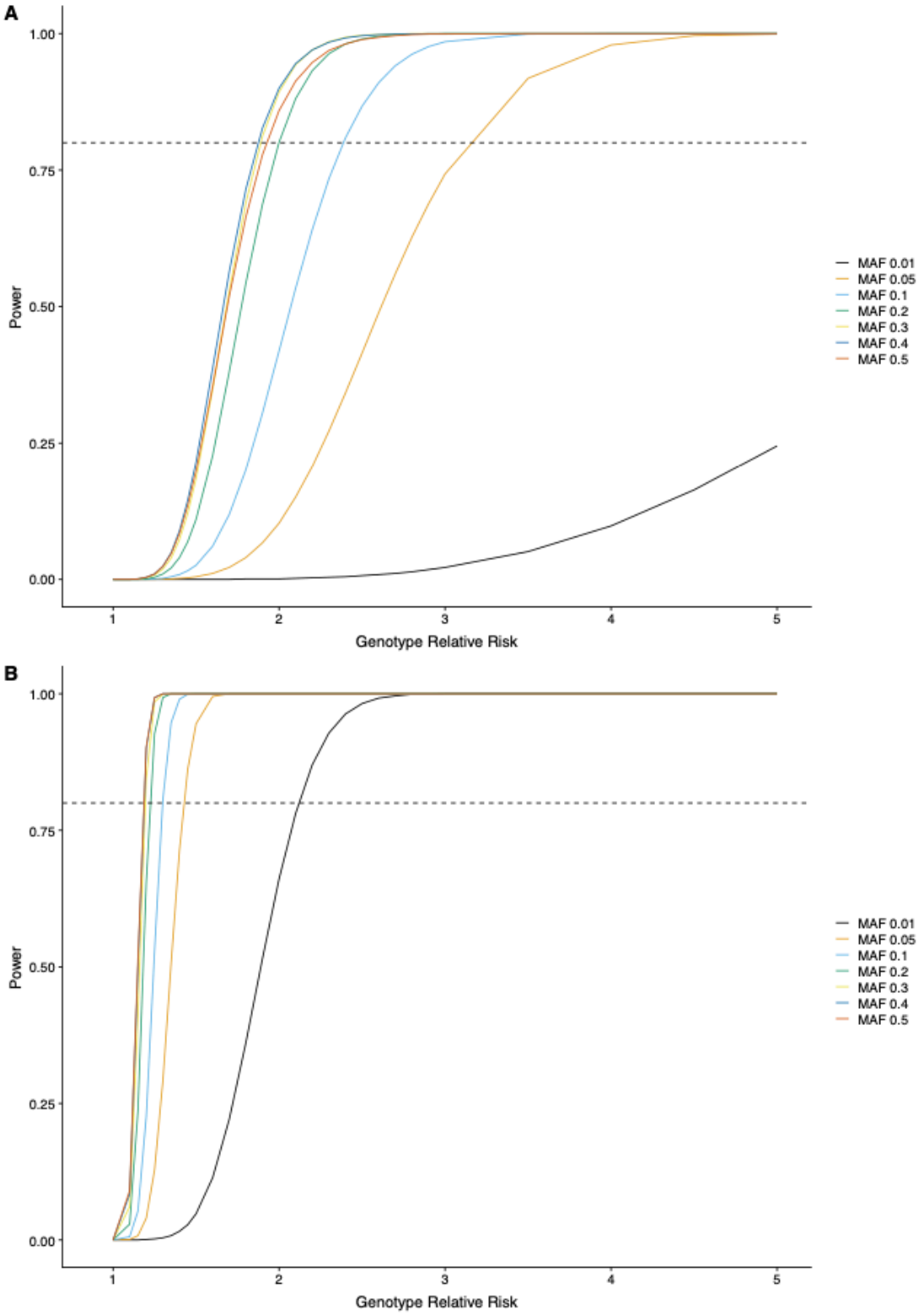


Figure 6.2: GWAS power calculations Panel A: Power calculation for BRRD case-control GWAS; Panel B: Power calculation for HCMR vs UKBB case-control GWAS

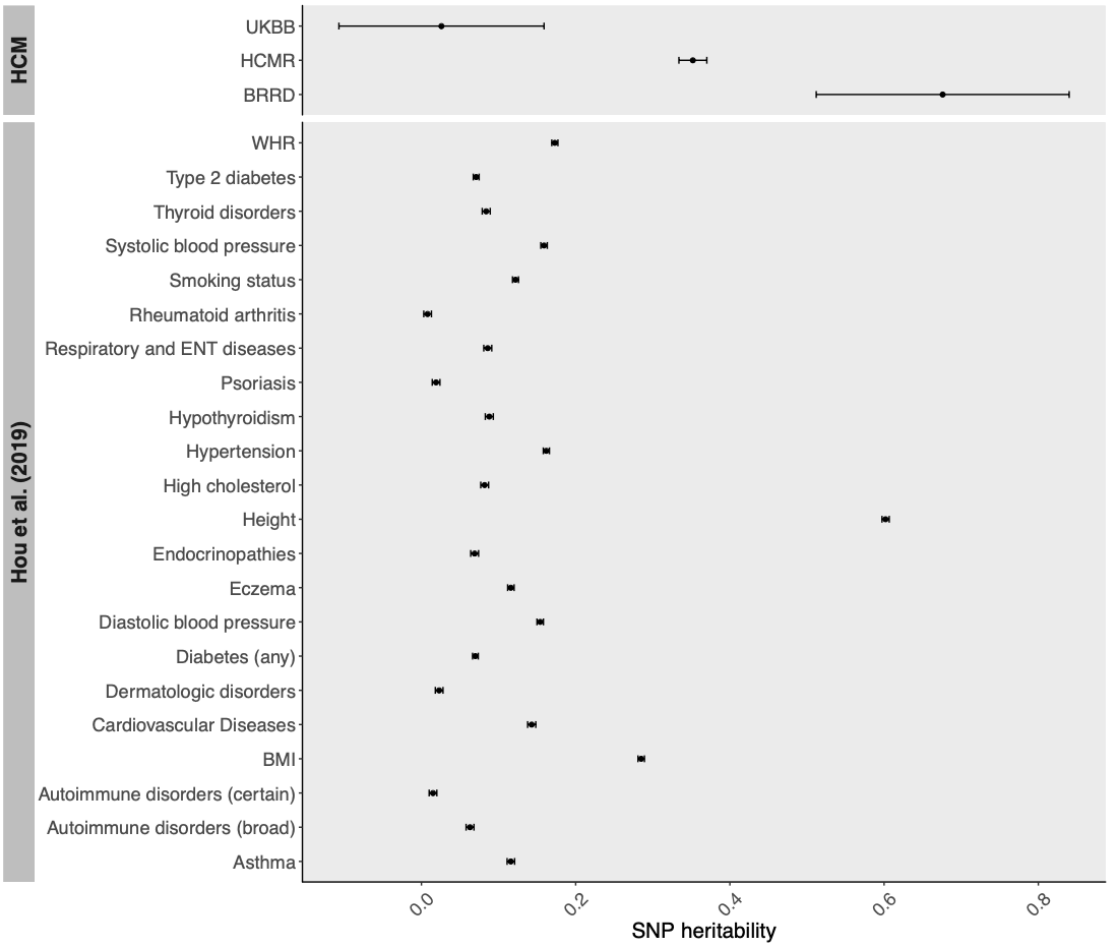


Figure 6.3: Comparison of SNP heritability estimates derived from HCM GWAS with other common traits and diseases SNP heritability estimates for a HCM and a range of common complex diseases generated by Hou et al. (2019) using a generalized random effects model (comparable to GREML-LDMS) in the UKBB.[301] Abbreviations: WHR: waist-hip ratio; BMI: body mass index; ENT: ear, nose and throat.

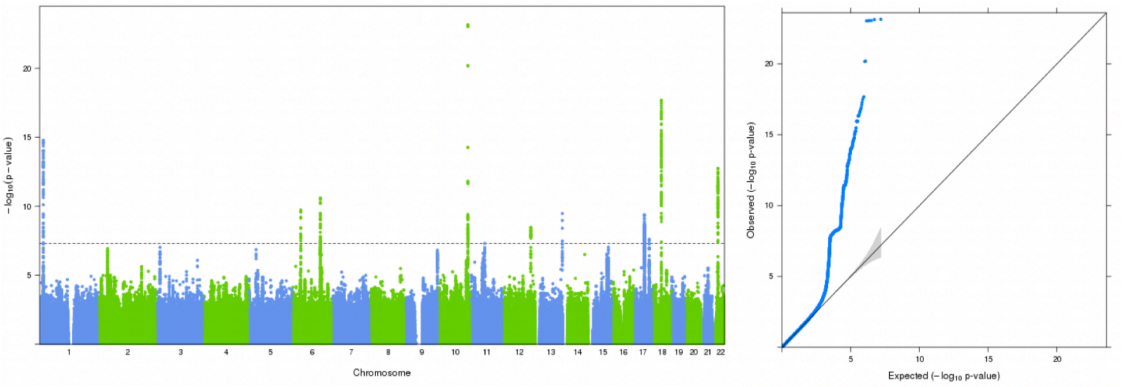


Figure 6.4: HCMR vs UKBB GWAS Multi-ancestry Manhattan plot and accompanying QQ plot. Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}). QQ plot of expected (x-axis) and observed (y-axis) -$\log_{10}$(p-value) values.

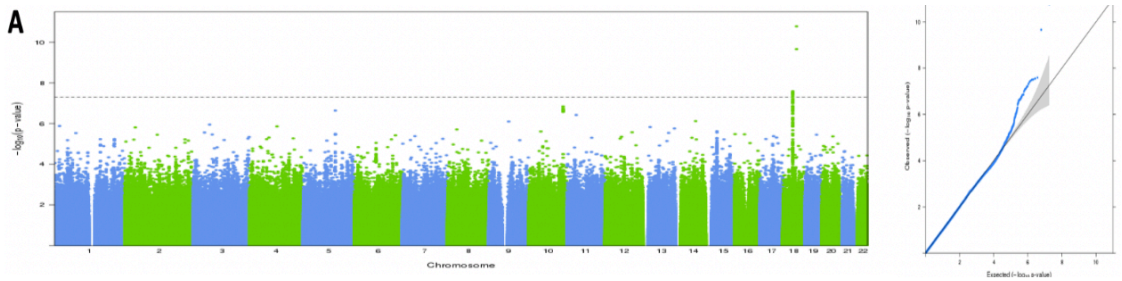


Figure 6.5: BRRD vs BRRD GWAS Multi-ancestry Manhattan plot and accompanying QQ plot. Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}). QQ plot of expected (x-axis) and observed (y-axis) -$\log_{10}$(p-value) values.

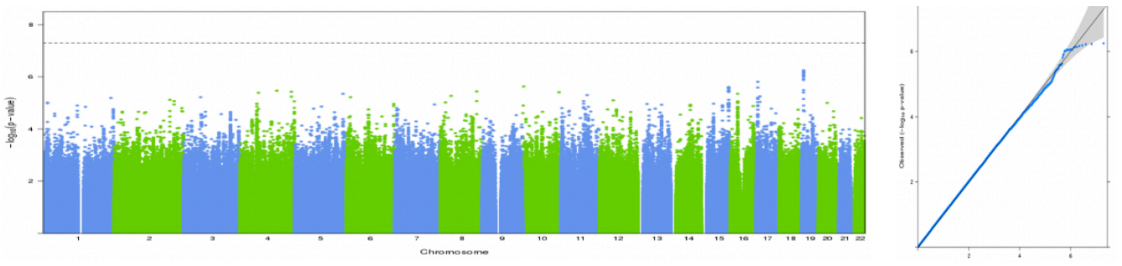


Figure 6.6: UKBB vs UKBB GWAS Multi-ancestry Manhattan plot and accompanying QQ plot. Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}). QQ plot of expected (x-axis) and observed (y-axis) -$\log_{10}$(p-value) values.

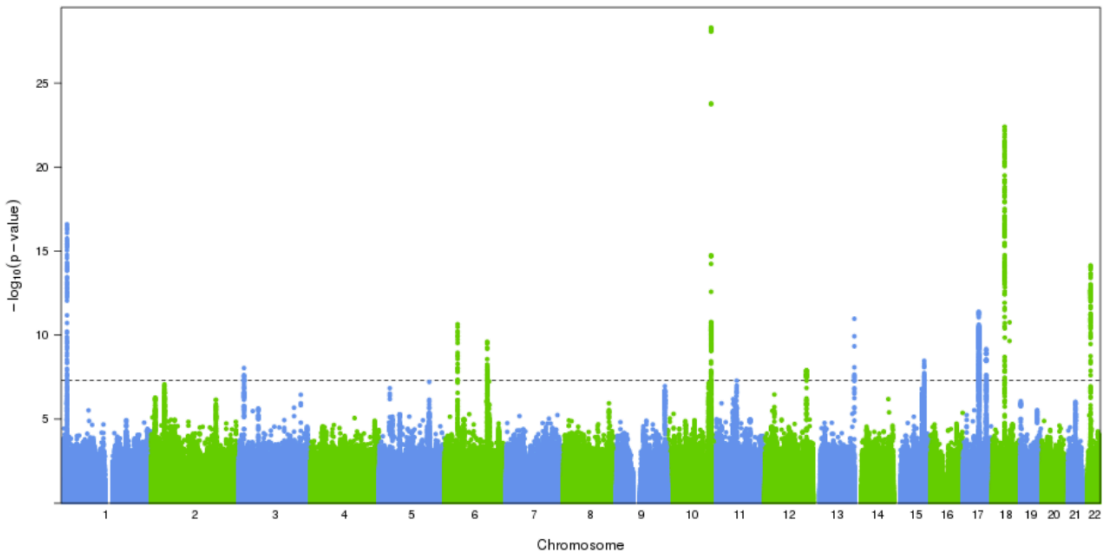


Figure 6.7: Multi-ancestry meta-analysis for HCMR vs UKBB and BRRD vs BRRD studies Multi-ancestry Manhattan plot and accompanying QQ plot. Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value = 5×10^{-8}). QQ plot of expected (x-axis) and observed (y-axis) -$\log_{10}$(p-value) values.

Chr	Start	Stop	SNPs contained within region	|Z| score		PPA model				
				Sarcomere negative	Sarcomere postive	0	1	2	3	4
1	5779370	6779370	rs846111	5.01	3.34	0.788	0.020	0.030	0.147	0.016
1	15840879	16858148	rs1048302, rs2015509	7.33	4.87	0.000	0.000	0.000	1.000	0.000
2	11099732	12099732	rs7556984	3.70	4.45	0.476	0.001	0.039	0.481	0.002
2	36623383	37623383	rs2003585	5.28	3.11	0.048	0.028	0.001	0.905	0.017
2	48351517	49351517	rs35258848	3.38	4.95	0.359	0.001	0.324	0.306	0.010
2	133277292	134277292	rs189654928	3.41	5.43	0.238	0.000	0.409	0.341	0.012
2	179221046	180221046	rs17362588, rs62177303	5.38	4.48	0.063	0.048	0.005	0.800	0.084
3	13797304	14797304	rs9843704, rs13061705	6.07	3.37	0.003	0.091	0.000	0.855	0.050
3	69406572	70406572	rs62253180	5.12	2.93	0.114	0.065	0.003	0.780	0.037
3	171300256	172300256	rs4894803	5.23	3.09	0.452	0.043	0.015	0.460	0.030
4	131347994	132347994	rs2312403	2.78	5.00	0.080	0.000	0.747	0.151	0.022
5	28926502	29926502	rs66761011	4.66	3.05	0.676	0.004	0.018	0.299	0.002
5	114121011	115121011	rs73780096	5.19	4.00	0.414	0.032	0.014	0.519	0.022
5	138168504	139168504	rs10052399	4.04	4.29	0.218	0.002	0.022	0.754	0.004
6	36147289	37147289	rs3176326	5.70	4.06	0.000	0.000	0.000	0.999	0.000
6	118164854	119206447	rs28436726, rs3890198, rs12212795	6.18	2.89	0.001	0.018	0.000	0.970	0.011
6	123318871	124318871	rs9320939	4.81	4.37	0.099	0.001	0.065	0.815	0.020
7	128049878	129049878	rs76600480	3.15	4.99	0.495	0.001	0.237	0.259	0.008
8	125351510	126351510	rs7003871	4.57	3.26	0.694	0.011	0.019	0.269	0.006
9	133989810	134989810	rs734638	4.14	3.81	0.045	0.001	0.007	0.945	0.002
10	114005037	115005037	rs11196085	4.72	4.46	0.119	0.004	0.009	0.861	0.006
10	120915685	121915685	rs72840788	10.83	4.51	0.000	0.000	0.000	1.000	0.000
11	70855	1516869	rs35469308	3.90	5.88	0.010	0.000	0.600	0.350	0.039
11	14428023	15428023	rs1390519	4.60	4.61	0.506	0.006	0.052	0.425	0.012
12	25845526	26845526	rs17380837, rs1480036	5.09	3.06	0.345	0.040	0.010	0.580	0.025
12	104158886	105158886	rs12299450	3.41	5.31	0.059	0.000	0.733	0.188	0.020
12	114862972	116106207	rs7977151, rs35357, rs7301677	5.15	3.72	0.007	0.002	0.000	0.987	0.003
13	29037036	30037036	rs2758215	4.04	5.60	0.016	0.000	0.774	0.174	0.036
13	113578558	115109853	rs41306688	5.90	4.06	0.000	0.000	0.000	1.000	0.000
15	78473393	79473393	rs8043123, rs1814880	5.22	3.95	0.077	0.022	0.004	0.872	0.025
15	84649575	85753258	rs748455, rs8033459	4.92	4.65	0.003	0.001	0.004	0.966	0.026
16	4787275	5787275	rs60247077	3.54	4.94	0.736	0.001	0.196	0.061	0.006
16	72259681	73259681	rs139235535	4.86	3.71	0.657	0.019	0.021	0.290	0.013
17	43188317	44488317	rs28768976	6.02	4.63	0.000	0.000	0.000	0.979	0.021
17	50864578	51864578	rs186910954	3.45	4.94	0.682	0.001	0.195	0.117	0.005
17	63807014	64807014	rs7210446	5.43	3.96	0.004	0.003	0.000	0.987	0.005
17	66822589	67822589	rs76905625	3.46	4.93	0.529	0.001	0.257	0.207	0.006
18	33744842	34780891	rs4799426, rs617207, rs118060942	10.37	3.71	0.000	0.014	0.000	0.969	0.018
18	47376108	48376263	rs142939703, rs1232572641	2.85	3.73	0.938	0.001	0.034	0.026	0.001
19	279606	1279606	rs117710064	4.74	3.89	0.670	0.007	0.026	0.291	0.006
19	45846330	46846330	rs8106955	4.96	3.62	0.345	0.053	0.012	0.553	0.037
21	30036712	31036712	rs2832230	4.07	4.32	0.168	0.003	0.044	0.768	0.017
22	23659307	24681652	rs6003909, rs2070458	9.01	3.50	0.000	0.114	0.000	0.786	0.100

Table 6.12: GWAS-PW results comparing sarcomere positive and sarcomere negative HCM Cells highlighted in red indicate the predominant model for associated loci. PPA model represents the posterior probability of association for each given model. Model 0: no association with either trait; Model 1: association with sarcomere negative, but not sarcomere positive; Model 2: association with sarcomere positive, but not sarcomere negative; Model 3: association with both sarcomere positive and sarcomere negative; Model 4: two independent signals in the same locus contribute towards both sarcomere positive and sarcomere negative separately.

SNP	EA	Sarcomere positive			Sarcomere negative			Power (p < 5e-8)	Power (5% FDR p < 1.6e-6)
		Beta	SE	*P*-value	Beta	SE	*P*-value		
rs72840788	G	-0.253	0.062	3.94E-05	-0.492	0.045	2.67E-27	99.45%	99.93%
rs617207	A	0.123	0.057	3.07E-02	0.43	0.041	3.66E-25	98.15%	99.69%
rs6003909	A	0.061	0.064	3.41E-01	0.406	0.045	2.20E-19	80.66%	93.56%
rs1048302	T	0.242	0.053	4.27E-06	0.291	0.04	2.46E-13	53.00%	76.67%
rs9843704	C	0.054	0.067	4.16E-01	0.321	0.053	1.31E-09	25.36%	49.59%
rs118060942	C	-0.22	0.225	3.29E-01	-1.036	0.132	4.25E-15	19.73%	42.15%
rs28768976	A	-0.225	0.06	1.59E-04	-0.268	0.045	1.82E-09	17.05%	38.25%
rs73780096	C	-0.157	0.101	1.19E-01	0.448	0.086	2.17E-07	15.79%	36.31%
rs41306688	A	-0.561	0.142	7.83E-05	-0.63	0.107	3.63E-09	15.51%	35.87%
rs28436726	G	-0.155	0.101	1.27E-01	-0.437	0.071	6.70E-10	12.80%	31.46%
rs3176326	G	-0.233	0.06	9.88E-05	-0.255	0.045	1.22E-08	11.87%	29.86%
rs7210446	G	-0.177	0.058	2.43E-03	-0.233	0.043	5.84E-08	7.32%	21.20%
rs2003585	T	0.1	0.052	5.62E-02	0.207	0.039	1.35E-07	6.94%	20.41%
rs846111	G	-0.028	0.065	6.67E-01	0.255	0.051	5.61E-07	6.69%	19.88%
rs17380837	C	0.061	0.059	2.98E-01	0.23	0.045	3.59E-07	6.44%	19.33%
rs7977151	A	-0.169	0.059	4.00E-03	-0.228	0.044	2.63E-07	5.89%	18.11%
rs4894803	A	-0.07	0.058	2.22E-01	-0.222	0.042	1.77E-07	5.41%	17.01%
rs8106955	T	-0.056	0.055	3.10E-01	-0.21	0.042	7.32E-07	5.13%	16.37%
rs8043123	C	-0.108	0.059	6.49E-02	-0.224	0.043	1.88E-07	5.13%	16.36%
rs748455	T	0.085	0.058	1.41E-01	0.22	0.045	8.90E-07	5.09%	16.28%
rs17362588	G	0.099	0.095	2.94E-01	-0.346	0.064	7.56E-08	3.67%	12.77%
rs9320939	G	-0.149	0.052	4.23E-03	-0.186	0.039	1.57E-06	2.98%	10.91%
rs62253180	A	0.1	0.072	1.66E-01	-0.255	0.05	3.11E-07	2.74%	10.25%
rs35357	A	0.063	0.072	3.86E-01	0.254	0.052	8.62E-07	2.65%	10.00%
rs139235535	C	-0.118	0.173	4.97E-01	-0.549	0.113	1.22E-06	1.13%	5.20%
rs3890198	T	0.023	0.051	6.56E-01	0.152	0.038	7.71E-05	0.64%	3.34%

Table 6.13: Evaluating discovery power in the sarcomere positive GWAS Discovery power, specific to disease-associated variants (<5% LFDR threshold) identified in the sarcomere negative HCM GWAS, was evaluated in the sarcomere positive HCM GWAS and demonstrated a systematically underpowered GWAS.

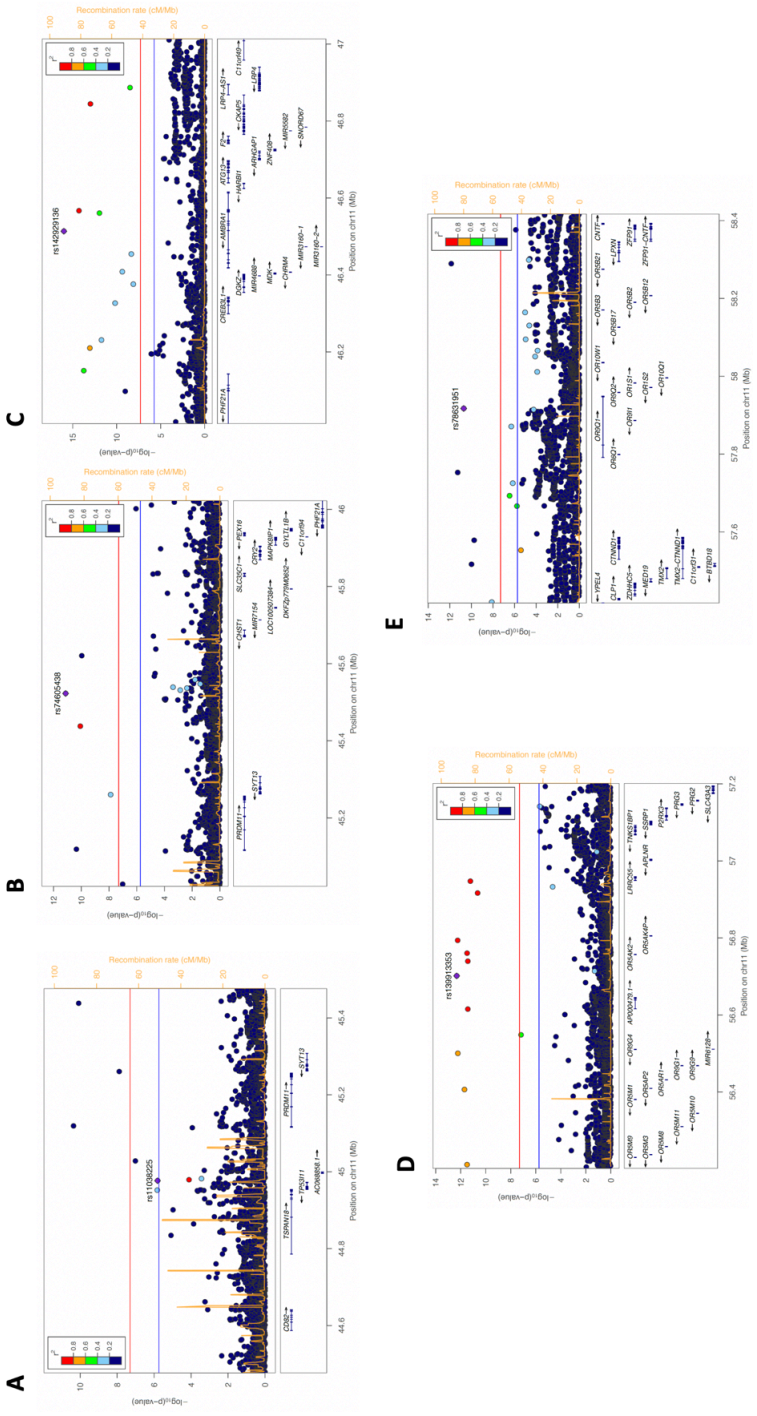


Figure 6.11: Regional association analysis for sarcomere positive HCMR vs UKBB GWAS. Variants identified via cojo analysis specific to the pericentromeric region of chromosome 11 are presented using LocusZoom. Panel A: rs11038225; panel B: rs74605438; panel C: rs142929136; panel D: rs139913353; panel E: rs78631951

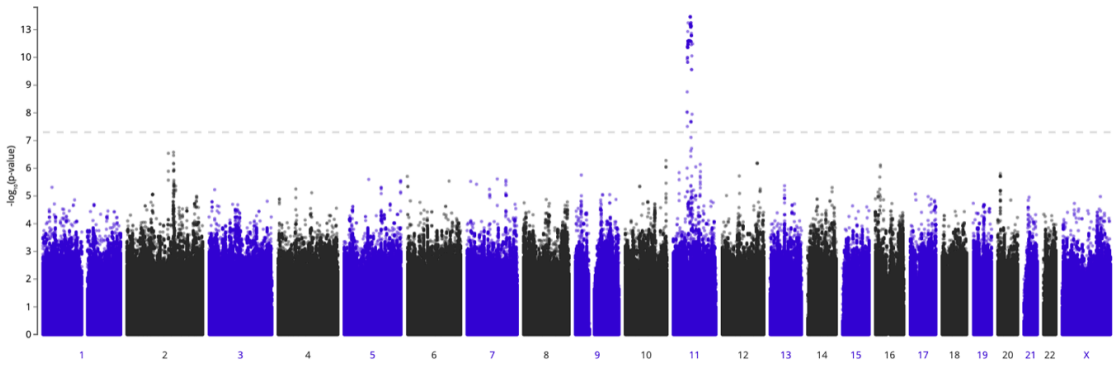


Figure 6.12: Manhattan plot for a HCM GWAS performed by the FINNGEN group Identification of a chromosome 11 region (40-60Mb) that demonstrates genome wide association (p-value $< 5\times10^{-8}$) in 192 cases and 436,274 controls from FINNGEN and the UKBB. Manhattan plot split by chromosome (x-axis) and -$\log_{10}$(p-value) (y-axis). Dotted line represents genome-wide significance threshold (p-value $= 5\times10^{-8}$).

Chr	Pos	REF	ALT	rsid	Nearest genes	MAF		OR (SE)	p-value
						Cases	Controls		
11	55423391	A	G	-	*OR4A15*	3.08E-02	3.91E-04	4.1e+25 (7.62)	9.9e-15
11	57380633	A	G	rs764568652	*PRG3*	3.06E-02	2.97E-04	5.6e+23 (7.25)	4.6e-14
11	48489468	C	T	rs140340264	*OR4A47*	2.99E-02	2.19E-04	1.9e+25 (7.72)	4.8e-14
11	56018565	T	TA	rs781774853	*OR5AS1*	3.08E-02	4.77E-04	1.6e+22 (6.83)	7.1e-14
11	58471197	G	A	rs546759453	*OR5B12*	2.96E-02	2.61E-04	1.5e+23 (7.17)	1.0e-13
11	47333566	G	A	rs397516005	*MYBPC3*	2.97E-02	2.49E-04	2.8e+20 (6.49)	4.0e-13
11	50787966	T	A	rs573034600	*OR4C12*	2.67E-02	2.67E-04	8.5e+21 (7.25)	3.3e-12
11	54837693	T	C	-	*OR4A5*	2.66E-02	2.72E-04	9.2e+21 (7.27)	3.4e-12
11	49125167	G	A	rs147205987	*FOLH1*	2.67E-02	2.57E-04	8.1e+21 (7.25)	3.5e-12
11	46339069	G	T	rs545277774	*DGKZ*	2.62E-02	3.12E-04	4.7e+19 (6.66)	1.1e-11
11	45691309	T	G	rs548817964	*CHST1*	2.21E-02	2.07E-04	3.00e+22 (8.03)	1.2e-10

Table 6.14: Summary level statistics from FINNGEN/UKBB HCM GWAS Results specific to variants demonstrating genome-wide significance (p-value $< 5\times10^{-8}$) from chromosome 11 (40-60Mb)

To evaluate whether common variant signals derived from the sarcomere positive HCMR vs UKBB GWAS are independent from rare pathogenic variants in *MYBPC3*, individual-level sequence and genotype data were combined, and haplotype analysis performed using a maximum likelihood method in Haploview (v4.2) (Figure 6.13).

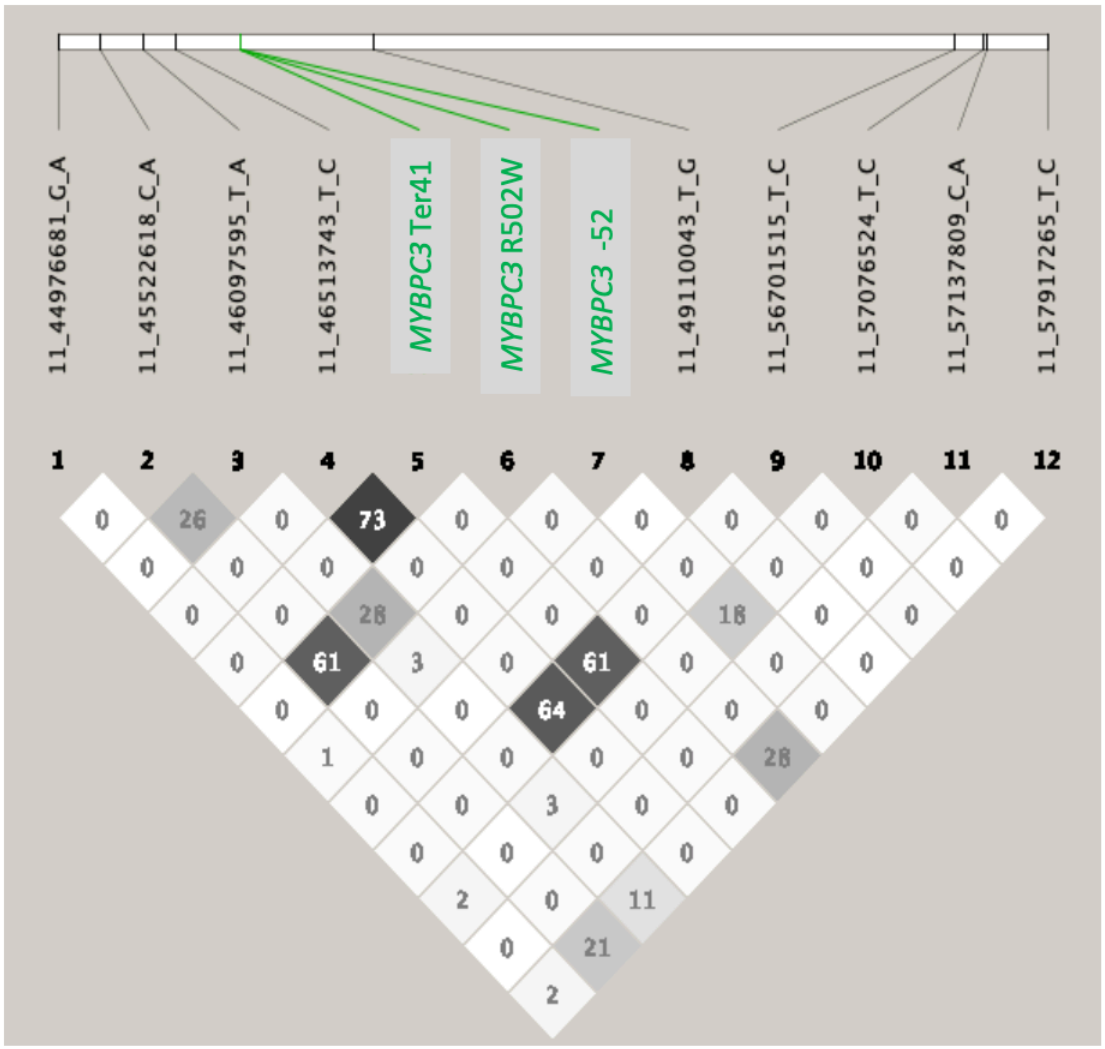


Figure 6.13: Haplotype analysis for the pericentromeric region of chromosome 11 r^2 values represented between common imputed variants (black) demonstrating genome-wide significance in the sarcomere positive GWAS and rare sequence variants (green)

Long range linkage disequilibrium between rare pathogenic variants in *MYBPC3* and common imputed variants in the neighbouring genomic regions was evaluated by combining gene panel sequence data and genotype array data for sarcomere positive individuals, from the HCMR cohort (n=844), and UKBB controls (n=19,612) in PLINK. Using HaploView (version 4.2), common imputed variants that demonstrated association, below the 5% LFDR threshold, with HCM were selected alongside well established and relatively frequent pathogenic variants in *MYBPC3* (specifically p.R502W, p.Trp792Valfs*41 (labelled Ter41 in Figure 6.13)

and c.1224-52G>A (labelled -52 in Figure 6.13) to allow for LD estimates to be calculated. *MYBPC3* p.R502W, *MYBPC3* p.Trp792Valfs*41 and *MYBPC3* c.1224-52G>A were selected for evaluation given that they are frequently detected pathogenic variants, with *MYBPC3* p.R502W and *MYBPC3* p.Trp792Valfs*41 previously demonstrating founder effects.

By implementing Firth's penalised logistic regression, it was possible to demonstrate that association signals observed for imputed common variants, that have been derived from the sarcomere positive GWAS, are synthetically associated with rare pathogenic variants in *MYBPC3* (Table 6.15). For example, *MYBPC3* p.Trp792Valfs*41 is in LD with 11_56701515_T_C (r^2=0.61) and when the effects each variant has on risk of HCM are simultaneously assessed, *MYBPC3* p.Trp792Valfs*41 appears significantly associated ($\beta = 6.82 \pm 1.48$), and 11_56701515_T_C demonstrates no effect towards HCM risk ($\beta = 0.214 \pm 0.402$). As a result, significantly associated SNPs reported within the 40Mb to 60Mb region of chromosome 11 were ignored in downstream analyses.

	Variant ID	r^2	Joint beta estimate (SE)
1	*MYBPC3* R502W	0.61	7.56 (1.49)
	11_45522618_C_A		-0.455 (0.398)
2	*MYBPC3* Ter41	0.61	6.82 (1.48)
	11_56701515_T_C		0.214 (0.402)
3	*MYBPC3* Ter41	0.73	7.66 (1.48)
	11_46513743_T_C		-0.344 (0.331)

Table 6.15: Joint effect of rare, pathogenic sequence variants and genome-wide common imputed variants upon HCM disease risk. Output from Firth's penalised logistic regression

6.5 Evaluate an individual's risk of developing HCM through the aggregate burden of common genetic variants

6.5.1 Constructing the genetic risk score instrument

All independent variants identified through the multi-ancestry meta-analysis were first evaluated for evidence of extreme pleiotropy using PhenoScanner.[347] rs28768976 demonstrated extreme pleiotropy (i.e. over 100 phenotypic assocaitions are listed), with a range of diverse traits that indicated that rs28768976 may be an unreliable marker for use in a genetic instrument. Therefore, after excluding rs28768976 for extreme pleiotropy, and rs78310129 due to concerns regarding long-range LD with pathogenic *MYBPC3* variants (see 6.4.8), 27 SNPs demonstrating independent associations with HCM at the 5% LFDR threshold in the all-comer HCM meta-analysis were aggregated for each individual using the allelic scoring function in PLINK. Weights were assigned from the beta estimates, generated from the multi-ancestry meta-analysis joint model GCTA-cojo results. Raw GRSs were plotted and evaluated, before being scaled to represent per-standard deviation effects (i.e. mean = 0, variance = 1) (Table 6.16).

6.5.2 Cohorts to evaluate a genetic risk score instrument

The GRS was used to evaluate risk of HCM in two discovery cohorts: 1. HCMR vs UKBB; and 2. BRRD; and in three independent validation cohorts: 1. Genomics England's 100,000 Genomes Project; 2. Royal Brompton Hospital's HCM case-control series; and 3. Amsterdam Medical Centre HCM case-control series.

Genomics England

Details of the Genomics England case series were provided by Dr Loukas Moutsianas (Genomics England). I had no involvement in either the recruitment or cohort-level quality control processing. Access to GeL was provided via the Cardiovascular GeCIP, via Registry ID RR254. GeL sequenced 101,162 genomes from 90,643 individuals recruited via 13 Genomic Medicine Centres across the United Kingdom.

rsID	Chr	Position	NEA	EA	Additive beta	Discovery		Validation		
						HCMR	BRRD	GeL	RBH	AMC
rs1048302	1	16340879	G	T	0.276737	x	x	x	x	x
rs7556984	2	11599732	A	G	0.185746	x	x	x	x	x
rs2003585	2	37123383	C	T	0.173631	x	x	x	x	x
rs62177303	2	179768624	T	C	0.174706	x	x	x	x	x
rs13061705	3	14291129	T	C	0.224098	x	x	x	x	x
rs4894803	3	171800256	G	A	-0.179213	x	x	x	x	x
rs66761011	5	29426502	G	A	-0.3486	x	x	NA	NA	x
rs10052399	5	138668504	C	T	0.206369	x	x	x	x	x
rs3176326	6	36647289	A	G	-0.246957	x	x	x	x	x
rs12212795	6	118654308	C	G	-0.392746	x	x	x	x	x
rs9320939	6	123818871	A	G	-0.173712	x	x	x	x	x
rs7003871	8	125851510	T	C	0.169008	x	x	x	x	x
rs734638	9	134489810	G	C	-0.186446	x	x	x	x	x
rs11196085	10	114505037	C	T	-0.189678	x	x	x	x	x
rs72840788	10	121415685	A	G	-0.420873	x	x	x	x	x
rs1390519	11	14928023	G	A	-0.2027	x	x	x	NA	x
rs1480036	12	26344726	C	T	-0.210931	x	x	x	x	x
rs7301677	12	115381147	T	C	0.213645	x	x	x	x	x
rs41306688	13	114078558	C	A	-0.601043	x	x	x	x	x
rs1814880	15	79021140	C	T	0.179413	x	x	x	x	x
rs8033459	15	85253258	T	C	-0.189921	x	x	x	x	x
rs7210446	17	64307014	A	G	-0.219756	x	x	x	x	x
rs118060942	18	34280732	T	C	-0.798616	x	x	x	x	x
rs4799426	18	34280891	G	A	-0.320945	x	x	x	x	x
rs117710064	19	779606	T	C	-0.222253	x	x	x	x	x
rs2832230	21	30536712	T	G	0.250812	x	x	x	x	x
rs2070458	22	24159307	T	A	0.292886	x	x	x	x	x

Table 6.16: Genetic variants and weights assigned to generate HCM genetic risk score

Analyses were performed on a subset of the total cohort, specifically 38,344 distantly related individuals, with genome sequence data mapped to GRCh38 that passed quality control criteria (including: $\geq$250 bp insert size, $\geq$75% mapped reads, $<$2% chimeric DNA fragments and $<$5% cross contamination). Phenotypic information was provided via HES and clinician entered human phenotype ontology terms. 435 HCM probands were available for analysis.

Amsterdam Medical Centre

Details of the Amsterdam Medical Centre case series were provided by Dr Connie Bezzina (Amsterdam University Medical Centre) for the purposes of replicatioon. I had no involvement in either the recruitment or processing of patient data. 999 cases were identified using current diagnostic criteria (left ventricular wall thickness $\geq$15mm, or $\geq$13mm in presence of family history) from cardiovascular genetics referral centres in the Netherlands (Amsterdam University Medical Center, Erasmus

Medical Center and the University Medical Center Groningen).[124] 2,117 controls were derived from a population cohort study from the Netherlands. Genotyping was performed using the Illumina Infinium BeadChip, Illumina OmniExpress and Global Screening Array. SNPs were mapped to GRCh37, and removed if: missingness rate >5%, HWE test p-value $<1\times10^{-6}$ for controls or p-value $<1\times10^{-10}$ for cases, or MAF <0.05. Individuals were excluded when: missingness exceeded 3%, inbreeding coefficient $\geq$0.1, genotype-phenotype sex mismatch existed, proportional identity by descent >0.05, or non-European ancestry was indicated by principal components analysis. Phasing (Eagle2) and imputation (Haplotype reference consortium (HRCr1.1) panel) was performed on the Michigan Imputation Server v.1.0.2. SNPs with MAF>0.01 and a Minimac R^2>0.5 were retained.

Royal Brompton Hospital

Details of the Royal Brompton case series were provided by Dr James Ware (Imperial College, London) for the purposes of replication. I had no involvement in either the recruitment or processing of patient data. 411 HCM cases were recruited from the Royal Brompton & Harefield Hospitals NHS Trust Cardiovascular Research Biobank. 1,211 controls, screened for evidence of HCM using cardiac imaging, were recruited from the UK Digital Heart Project.[125] Genotyping was performed using the Illumina Human OmniExpress beadchip. SNPs were mapped to GRCh37 and excluded if MAF<0.01, HWE P $<1\times10^{-7}$, or missingness rate >0.05. Sample QC excluded samples with a genotype-phenotype sex mismatch, heterozygosity rate >3 standard deviations from the mean, missingness rate >0.03 or evidence of non-European ancestry via principal components. Genotypes were phased using SHAPEIT (v2.r790) and imputed using IMPUTE2 (v2.3.2), against the UK10K and 1000 Genomes Project reference panel. SNPs with MAF >0.01 and INFO score >0.4 were retained.

				Total (n)	Age (SD)	Gender (%)	sGRS (mean (SD))
Discovery	HCMR	All	Cases	2541	49.6 (11.2)	72.2	0.626 (1.03)
			Controls	40283	54.0 (6.95)	71.5	-0.04 (0.985)
		S-	Cases	1635	51.4 (10.4)	75.8	0.718 (1.04)
			Controls	20141	53.9 (6.92)	71.9	-0.039 (0.983)
		S+	Cases	871	46.2 (12.0)	65.9	0.455 (0.985)
			Controls	20142	54.1 (6.97)	71.2	-0.04 (0.987)
	BRRD	All/S-	Cases	233	57.8 (11.1)	80.4	0.795 (1.01)
			Controls	4680	45.7 (20.1)	40.2	-0.033 (0.986)
Validation	GeL	All	Cases	435	53.5 (15.0)	69.5	0.569 (1.09)
			Controls	36500	46.3 (20.2)	45.8	-0.015 (0.996)
		S-	Cases	371	54.3 (14.7)	73.3	0.616 (1.08)
			Controls	18250	46.3 (19.8)	46.3	-0.01 (1)
		S+	Cases	64	49.0 (15.7)	51.6	0.302 (1.11)
			Controls	18250	46.4 (20.7)	45.3	-0.015 (0.997)
	RBH	All	Cases	411	65.7 (15.1)	71	0.4 (1.06)
			Controls	1211	47.0 (13.3)	53.9	-0.136 (0.941)
		S-	Cases	316	68.3 (14.1)	72.5	0.452 (1.07)
			Controls	606	46.8 (13.4)	45.9	-0.151 (0.946)
		S+	Cases	95	56.9 (14.8)	66.3	0.229 (1.03)
			Controls	605	47.2 (13.3)	53.7	-0.12 (0.936)
	AMC	All	Cases	999	NA	65.6	0.358 (1.01)
			Controls	2117	NA	56.4	-0.169 (0.951)
		S-	Cases	552	NA	63.0	0.467 (1.02)
			Controls	1059	NA	56.6	-0.136 (0.962)
		S+	Cases	446	NA	68.8	0.227 (0.977)
			Controls	1058	NA	56.4	-0.202 (0.939)

Table 6.17: Summary of cohorts included in GRS analysis Abbreviations: All: all-comers; S-: sarcomere negative; S+: sarcomere positive; AMC: Amsterdam Medical Centre; RBH: Royal Brompton Hospital; GeL: Genomics England 100,000 Genomes project

6.5.3 Evaluation of a genetic risk score instrument

A logistic regression model was fitted with affected status as the outcome variable and standardised GRS score as an explanatory variable with covariates including age, gender, and the first ten ancestrally informative principal components (Table 6.17). Effect sizes were estimated in terms of per-standard deviation effects and also using a quintile-based analysis to assess the top and bottom 20% of the population, with the central 60% acting as a reference group. Fixed-effects meta-analysis was implemented to assess the overall effect of the GRS in the discovery and validation cohorts, for the entire HCM population, and also stratified by sarcomere carrier status.

6.5.4 HCM genetic risk score performance

All comer analysis

Using a per-standard deviation effect size, the GRS appears to predict HCM risk more strongly in the discovery cohort (OR=1.99 [95% CI:1.91-2.08]) than in the validation meta-analysis (OR=1.75 [95% CI:1.65-1.85]) (Figure 6.14). This is consistent with the "winners curse" phenomenon, an ascertainment bias that results in the true genetic effect size of a variant being lower than that reported in an initial discovery cohort. [348] Stratification of the HCMR vs UKBB cohort by their average genetic ancestry, as determined by PCA, demonstrates similar effect sizes across all ancestry groups (Table 6.18). Individuals in the lowest quintile appear protected from HCM (discovery: OR=0.37 [95% CI:0.31-0.43]; validation OR=0.51 [95% CI:0.43-0.61]), whereas individuals in the top 20% demonstrate a greater than a two-fold increased risk (discovery OR=2.66 [95% CI:2.44–2.91]); validation OR=2.33 [95% CI:2.05–2.65]).

Sarcomere-status analysis

When partitioned by sarcomere carrier status, per-standard deviation effects were larger in the sarcomere-negative GRS analysis (OR=2.04 [95% CI:1.96-2.13]) than the sarcomere-positive GRS analysis (OR=1.62 [95% CI:1.52-1.71]), in alignment with ${h^2}_g$ estimates (Figures 6.15 and 6.16). Quintile based analyses supported this observation, with the GRS conferring more extreme effect sizes in the sarcomere-negative cohort than the sarcomere-positive cohort.

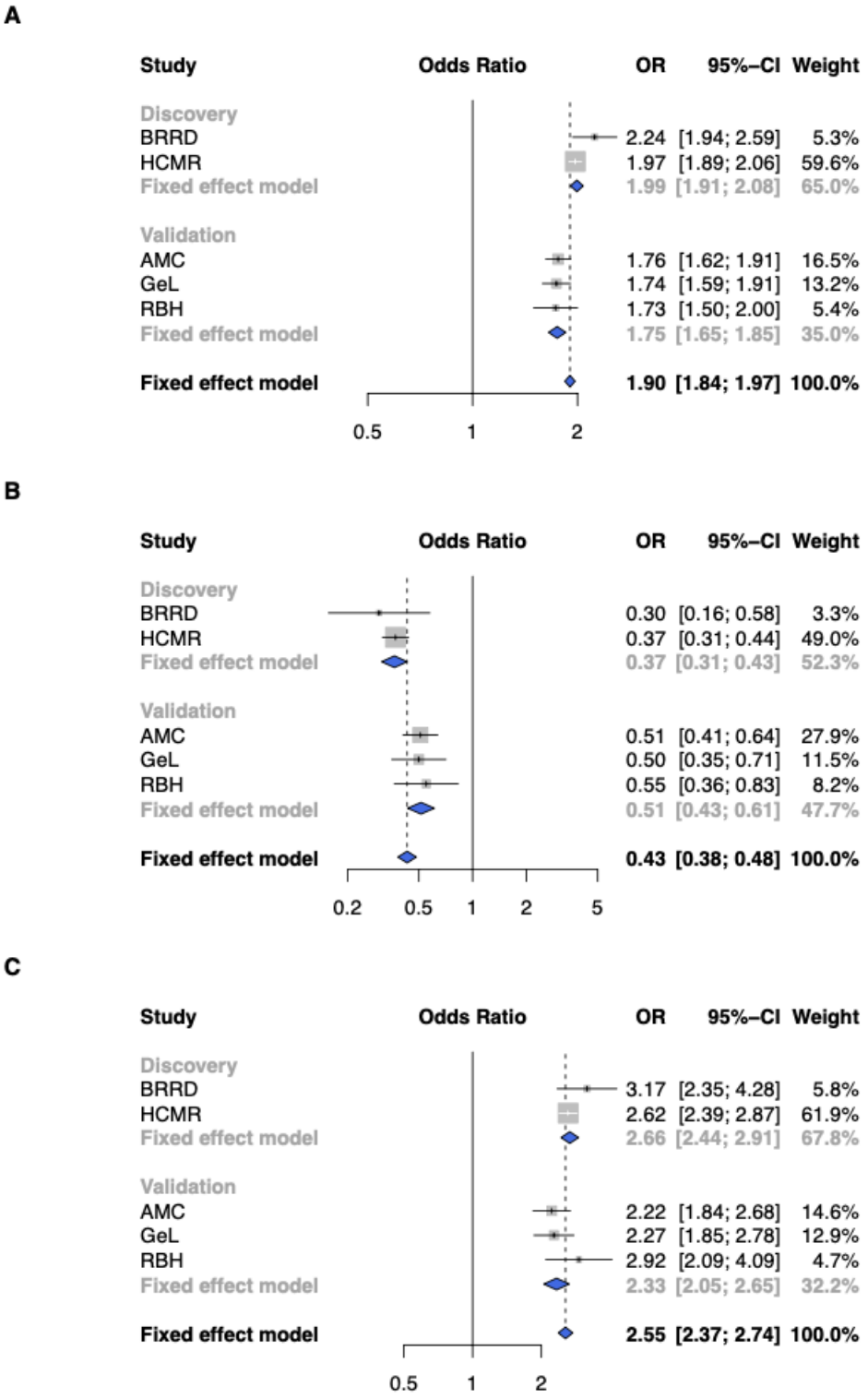


Figure 6.14: Forest plot assessing the performance of a HCM genetic risk score in all comers. HCMR: HCM registry; GeL: Genomics England 100,000 Genomes; RBH: Royal Brompton Hospital cohort; AMC: Amsterdam Medical Center. Panel A) Per standard deviation analysis; Panel B) Effect size corresponding to the bottom 20% of the population compared with the middle 60% of the population in a quintile based analysis; Panel C) Effect size corresponding to the top 20% of the population compared with the middle 60% of the population in a quintile based analysis.

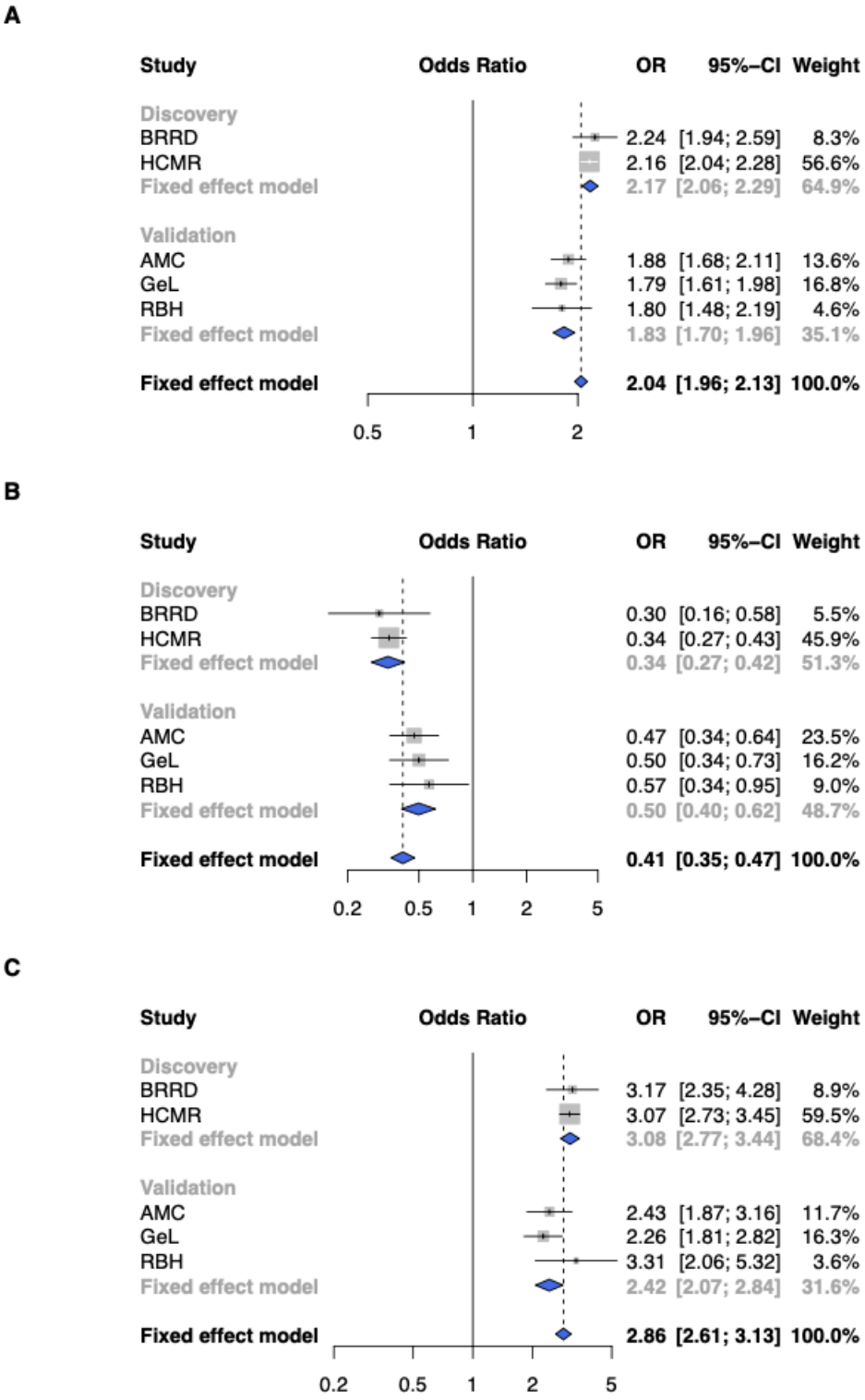


Figure 6.15: Forest plot assessing the performance of a HCM genetic risk score in sarcomere negative HCM HCMR: HCM registry; GeL: Genomics England 100,000 Genomes; RBH: Royal Brompton Hospital cohort; AMC: Amsterdam Medical Center. Panel A) Per standard deviation analysis; Panel B) Effect size corresponding to the bottom 20% of the population compared with the middle 60% of the population in a quintile based analysis; Panel C) Effect size corresponding to the top 20% of the population compared with the middle 60% of the population in a quintile based analysis.

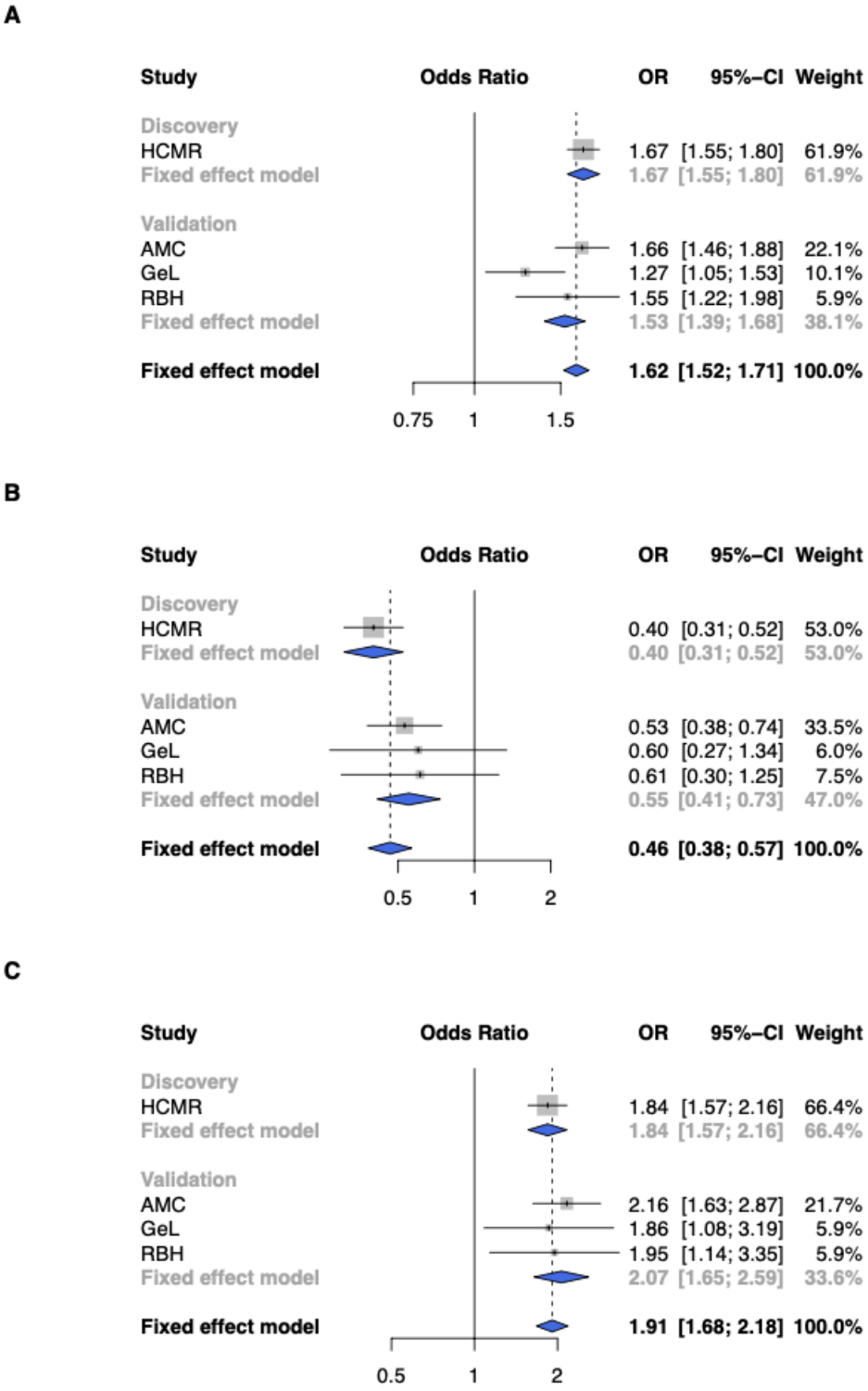


Figure 6.16: Forest plot assessing the performance of a HCM genetic risk score in sarcomere positive HCM HCMR: HCM registry; GeL: Genomics England 100,000 Genomes; RBH: Royal Brompton Hospital cohort; AMC: Amsterdam Medical Center. Panel A) Per standard deviation analysis; Panel B) Effect size corresponding to the bottom 20% of the population compared with the middle 60% of the population in a quintile based analysis; Panel C) Effect size corresponding to the top 20% of the population compared with the middle 60% of the population in a quintile based analysis.

Ancestry	Sarcomere status	N	sGRS		OR [95% CI] per SD unit	p-value
			Mean	SD		
AFR	Control	1148	0.089	0.771	2.04 [1.65-2.53]	8.15E-11
	Sarcomere -ve	145	0.519	0.795		
	Sarcomere +ve	49	0.465	0.656		
AMR	Control	170	-0.248	0.964	1.74 [1.18-2.62]	6.27E-03
	Sarcomere -ve	33	0.354	1.068		
	Sarcomere +ve	11	0.671	1.023		
EAS	Control	267	-0.35	0.731	2.00 [1.35-3.02]	6.57E-04
	Sarcomere -ve	46	0.084	0.796		
	Sarcomere +ve	17	-0.096	0.588		
EUR	Control	37203	-0.053	0.991	1.98 [1.88-2.07]	7.06E-169
	Sarcomere -ve	1284	0.73	1.047		
	Sarcomere +ve	729	0.486	1.003		
FIN	Control	288	-0.154	0.937	2.46 [1.76-3.55]	3.79E-07
	Sarcomere -ve	39	0.833	1.116		
	Sarcomere +ve	22	0.768	1.21		
SAS	Control	1207	0.356	0.949	2.20 [1.78-2.74]	5.75E-13
	Sarcomere -ve	88	1.156	0.926		
	Sarcomere +ve	42	0.796	1.003		

Table 6.18: Genetic risk score stratified by ancestry, as determined via principal components analysis in the HCMR vs UKBB cohort. Abbreviations: AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian

6.6 Discussion and limitations

Evaluating common genetic variants across the genome in individuals with and without HCM has provided evidence to support the hypothesis that common variants contribute towards the genetic architecture of HCM. This finding is a departure from the traditional and largely reductionist approach to rare diseases, that assumed disease was attributable to a single, highly penetrant genetic variant. The contribution of common variants towards HCM is surprisingly high, as demonstrated by the relatively large h^2_g estimates, and exceed those reported for other cardiovascular diseases such as coronary artery disease.

The meta-analysis of two discovery GWAS identified 28 discrete loci, all but one of which are novel. The *FHOD3* locus has previously been reported and is replicated in this analysis.[118] Whilst many of the loci identified through these GWAS are of unknown function, there are several genes nested within regions of

associations that could plausibly influence susceptibility towards HCM, including *BAG3, FHOD3, HSPB7, ADPRHL1* and *SLC6A6*. This is a general limitation of genome-wide association studies and further analysis is required before the mechanism underpinning disease risk for each variant can be comprehended.[319] A further limitation relates to the overall sample size. Future studies that enrol larger case cohorts will be capable of reporting, with greater precision, the risk estimates associated with each susceptibility variant and the aggregate effects. Here, in aggregate the common genetic variants identified through these GWAS efforts contribute towards HCM risk-susceptibility: individuals with a GRS in the top 20% of the population are at a greater than a two-fold increased risk of developing HCM than those in the central 60% of the population.

Whilst the genetic architecture of sarcomere positive and sarcomere negative HCM appear to be highly genetically correlated, ${h^2}_{\mathrm{g}}$ and GRS analyses indicate that common variants confer larger effects in the sarcomere negative cohort. Future studies that incorporate larger case-series, particularly sarcomere positive cases, will be able to refine these findings. Presently, these findings could partially explain why the majority of individuals with HCM (i.e. those without a disease-causing sarcomere variant) develop HCM. Furthermore, subsequent analysis could be targeted towards assessing whether heritable risk factors, commonly detected in the sarcomere negative population, such as hypertension or obesity, influence an individual's risk of developing HCM using two-sample Mendelian randomisation.[56, 340, 349] Two-sample Mendelian randomisation (2SMR) is an extension to Mendelian randomisation, a statistical methodology that uses genetic variation to infer causal relationships between exposures (i.e. risk factors measured using SNP-exposure effects) and outcomes (i.e. disease risk measured using SNP-outcome effects). Traditionally the influence a SNP had on an exposure and a outcome were inferred from a single set of samples. [350] However, 2SMR facilitates SNP-exposure and SNP-outcome effects to be derived from separate samples. This has enabled pre-existing summary statistics from GWAS to be leveraged for the purposes of evaluating causality between an exposure and a diverse range of outcomes.[351]

7 Conclusions

Contents

7.1 Introduction

The purpose of this thesis was to explore the broad hypothesis that the genetic architecture of HCM extends beyond the rare variant contributions that have been extensively investigated to date. To tackle this hypothesis four key objectives were defined, specifically: 1) evaluate monogenic disease models of HCM (Chapter 3); 2) consider possible oligogenic causes of HCM (Chapter 4); 3) assess the penetrance associated with disease causing variants in HCM (Chapter 5); and 4) investigate

the role of common variants in HCM (Chapter 6).

Here, I will provide further discussion regarding each of these objectives, consider whether sufficient evidence has been generated to reject the null hypothesis, and discuss the clinical implications.

7.2 Monogenic disease models of HCM

By studying the monogenic disease model of HCM, the null hypothesis (i.e. that the genetic architecture of HCM does not extend beyond rare variant contributions) was formally defined using data that would later be used to explore alternative genetic architecture hypotheses.

As presented in Chapter 3, the null hypothesis was appropriately defined using a relatively large case-control study, composed of individual-level DNA sequence data from ∼5,400 HCM cases and ∼12,300 controls. The proportion of HCM cases harbouring a pathogenic or likely pathogenic variant in at least one of eight core sarcomere genes (*MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3, TPM1* or *ACTC1*) across the OMGL and HCMR cohorts (25.9% [95% CI: 24.7-27.1%] (n=1396/5,393)) was slightly lower than perhaps anticipated when compared with other historic cohorts (Walsh et al. (2016) and Alfares et al. (2016) both report a pathogenic/likely pathogenic yield of ∼32%), but this could potentially reflect advances in variant classification methodologies.[8, 55, 91]

Evaluation of the OMGL-HCMR cohort has confirmed many previous findings. This includes burden analysis, which reproduced gene-level findings from Walsh et al.(2016), and provided further evidence supporting the role of *CSRP3* and *FLNC* in HCM. [8, 60, 216] Variant-level findings were also replicated with the OMGL-HCMR cohort, which provided supporting evidence that *MYBPC3* p.Arg502Trp is the most frequently observed pathogenic variant in HCM. However, haplotype analysis of *MYBPC3* p.Arg502Trp was not able to convincingly demonstrate the founder effects that have been previously reported.[8, 179, 263]

Analysis of the OMGL-HCMR cohort has provided evidence to challenge previous assumptions. For example, prior reports had quoted up to ∼8% of HCM cohorts

would yield individuals with multiple pathogenic/likely pathogenic variants. The OMGL-HCMR cohort aligns with more contemporary data indicating this presentation is rare, and more likely to effect ∼1:1000 HCM cases.[176] Additionally, the OMGL-HCMR data provides supporting evidence for clinical diagnostic laboratories to extend their genomic capture regions so that non-coding regions, particularly across *MYBPC3* where cryptic splice sites can lead to haploinsufficiency, can be surveyed.[352] Haploinsufficiency is a relatively common disease mechanism that refers to the intolerance of loss-of-function variants in diploid organisms. The National Cancer Institute defines haploinsufficieny as: "the situation that occurs when one copy of a gene is inactivated or deleted and the remaining functional copy of the gene is not adequate to produce the needed gene product to preserve normal function".[353, 354] A key finding that supports this recommendation was the discovery that *MYBPC3* c.1224-52G>A, a deep intronic variant that introduces a cryptic splice acceptor site and leads to a premature termination codon (p.Ser408fs*31), accounts for ∼1% of HCM cases. Furthermore, it was proven that *MYBPC3* c.1224-52G>A frequently occurs in LD with *MYBPC3*Δ25, a variant previously associated with HCM despite its allele frequency being incongruent with the prevalence of HCM, and consequently explains HCM disease risk in a small proportion of the ∼100 million individuals who possess *MYBPC3*Δ25 worldwide (as presented in Chapter 4.[180]

Collectively, exploration into the monogenic architecture of HCM has convincingly characterised the null hypothesis and has provided actionable information that can be used to improve the care provided to patients and families affected by HCM. However, it is striking that a monogenic hypothesis only explains ∼26% of HCM (i.e. those with a pathogenic or likely pathogenic variant), with the genetic aetiology for the vast majority (∼74%) of HCM yet to be explained.

7.3 Oligogenic disease models of HCM

Oligogenicity represents a mode of genetic inheritance that has often been hypothesised to contribute towards the genetic architecture of disease, but very few

examples exist.[104, 106] Having accounted for monogenic causes of HCM in Chapter 3, Chapter 4 intended to explore oligogenicity in a sarcomere negative cohort. The European subset of BRRD cohort, as determined by PCA, was selected to perform an exploratory analysis given that BRRD cases and controls underwent genome sequencing and bioinformatic processing in unison, which limited a potential source of technical bias, potentially present in other cohorts. However, the relatively small HCM cohort limited insights that could be gleaned with respect to carriage of multiple rare (i.e. MAF <0.01) variants across these 8 sarcomere genes. Power calculations using data from the BRRD, regarding the proportion of controls (n=5,685) harbouring multiple rare variants across *MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3, TPM1* and *ACTC1*, indicates at least a five-fold increase in the number of sarcomere-negative HCM cases would be required to generate 80% experimental power, assuming a relative risk of at least 3, and 1% cases and 0.4% of controls harbour at least two protein-altering variants (MAF <0.001). Insufficient data are currently available to formally assess this hypothesis.

Whilst a systematic assessment of oligogenicity in HCM was underpowered using the BRRD, our understanding of oligogenicity was advanced through a conditional analysis of a candidate variant, *MYBPC3*Δ25, that had been previously posited to be a cornerstone of oligogencity in HCM. As discussed previously in section 7.2, analysis performed using OMGL-HCMR data refutes this assertion.

7.4 Penetrance

Whilst HCM is underpinned by rare variants in core sarcomere genes, many pathogenic and likely pathogenic variants have been observed in individuals without a personal or family history of disease, and have become known as secondary findings (SFs). This observation is important, particularly as genome sequencing is increasingly adopted for diagnostic purposes across a diverse range of rare disease and cancers, and an increasing number of individuals learn that they carry SFs.[355] The provision of health care services to accommodate SF carriers is contingent on better understanding SFs, both in terms of the prevalence and clinical impact.

In Chapter 5, the prevalence of SFs, relating specifically to HCM, was approximated in both the BRRD (0.28% [95% CI:0.18 - 0.43%]) and in gnomAD/TOPMed (0.31% [95% CI: 0.28 – 0.33]). These estimates have benefited from contemporary variant classifications, informed by the availability of large reference datasets such as gnomAD, which makes direct comparison with studies conducted prior to this era challenging. For instance, analysis conducted in 2012 suggested the prevalence of SFs in the Framingham Heart Study (FHS) and Jackson Heart Study (JHS) was 0.6%, approximately two-fold higher than what is observed in the BRRD and gnomAD/TOPMed.[279] Contemporary analysis performed in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort provided an estimate for HCM genes (0.13% [95%CI:0.05-0.29%]) that overlaps with those derived from BRRD and gnomAD/TOPMed.[280]

However, the probability an individual will develop HCM, conditioned on the presence of a disease causing variant (i.e. the penetrance) remains poorly understood. Part of this is attributable to the genetic architecture of HCM, a relatively rare disease ($\sim$1/500) characterised by allelic heterogeneity, but also to methodological challenges, both in terms of avoiding ascertainment bias and in assigning affected status to individuals.

It had been envisaged that a genotype-first approach in a large, longitudinal biobank with accurate phenotypic recording would circumvent many of these prior challenges. However, analysis performed in Chapters 5 and 6 using the UKBB raised concerns regarding the validity of HCM diagnostic codes assigned to individuals and further analysis is required before the UKBB can be used to assess penetrance estimates. Nevertheless, a positive likelihood framework was designed to approximate SF penetrance using available case-control data from the OMGL-HCMR series and gnomAD/TOPMed. It was hoped that through application of this positive likelihood framework, utilising >5,000 cases and >180,000 controls penetrance estimates with relatively narrow confidence intervals could be generated that may help inform clinical decision making. However, the high degree of allelic heterogeneity resulted in relatively few variants having sufficient counts to generate

precise penetrance estimates. Even for the four variants that were present in at least 10 cases and 10 controls, penetrance estimates were relatively imprecise (see Table 5.2). This included *MYBPC3* p.Arg502Trp, which appears to have a penetrance of 36% [95% CI: 24 - 50%], similar to the 0.5 estimate that is widely quoted based on family-based studies[94, 179, 263].

Similar challenges were faced when attempting to quantify the clinical impact of SFs through a recruitment-by-genotype strategy. Despite an approach that took into consideration potential ethical issues, including violation of an individual's "right not to know", the clinical study reported in Chapter 5 provided only a handful of anecdotal reports and as such few meaningful conclusions emerged.[278] Whilst not designed for quantitative analysis, this recruitment-by-genotype study highlighted the recruitment challenges one may face with respect to recruiting sufficient individuals. The identification of a 70 year old carrier of *MYBPC3* p.Trp792fs, without a personal or family history of HCM, provides anecdotal evidence regarding genetic resilience.[281]

7.5 Polygenic disease models of HCM

The concept of polygenicity had traditionally been reserved for common complex diseases, such as type two diabetes or multiple sclerosis, however, analysis performed in Chapter 6 indicates common variants also influence susceptibility towards HCM. Heritability estimates performed using the HCMR and BRRD cohorts indicate that a larger proportion of HCM phenotypic variability is accounted for by the additive effects of common variants than in other widely studied cardiovascular diseases, such as coronary artery disease. It is acknowledged that there is an unknown amount of "missing" SNP heritability in HCM, whereas twin studies have been able to help quantify this "missing" heritability in coronary artery disease.[302]

Common variants appear to influence risk of HCM more for sarcomere negative HCM than in sarcomere positive HCM. However, it is apparent that disease risk in the sarcomere positive population is primarily driven by the presence of a disease-causing variant in a core sarcomere gene. Whether the aggregate burden of

common variants in sarcomere positive HCM acts to modify expressivity, similar to observations in other rare diseases, remains to be proven.[284–286]

Meta-analysis of two discovery HCM GWAS identified 28 discrete loci. All these loci were novel, with only the *FHOD3* locus being previously reported by Wooten et al (2013).[118] Of these associated loci, few are recognisable as established HCM genes. Instead attention is directed towards potential new disease mechanisms, through genes located within regions of association, such as *BAG3, FHOD3, HSPB7, ADPRHL1* and *SLC6A6*. Many of these loci have also been associated with DCM, and as anticipated from experiments evaluating contractility at a myofilament level, these loci appear to have opposing effects with respect to HCM and DCM phenotypes.[209]

7.6 Future work

This thesis presents a body of work that has aimed to evaluate the various genetic architecture models that may underpin HCM. However, it is evident that future work can be performed to further characterise monogenic, oligogenic and polygenic contributions. The discovery that the additive effects of common genetic variants have a relatively large effect on HCM disease risk is perhaps the most important finding from this body of research. Consequently, it is foreseeable that future efforts will be directed towards the recruitment of additional HCM case series in an attempt to generate additional discovery power so that meta-analyses can refine effect size estimates and identify additional loci. However, in the short term, data generated from these initial GWAS efforts are suitable for other experiments, with three areas that I believe should be prioritised, specifically: 1) evaluating the role of heritable risk factors in the aetiology of HCM; 2) to evaluate the role common variants, associated with HCM, have on the expressivity and/or penetrance of HCM; and 3) to help enrich gene-discovery efforts for those individuals likely to possess a monogenic form of disease. In the longer term, the results presented here could provide future direction for the development therapeutics specific to HCM, or potentially a more general heart failure population. However, moving

from the identification of novel susceptibility loci, to understanding the functional consequences of these variants, and then onto the development of a therapeutic, whilst challenging, should remain a longer term goal

7.6.1 Heritable risk factors and risk of HCM

First, to evaluate whether common heritable risk factors have a role in the aetiology of HCM, particularly sarcomere negative HCM. For instance, I would hypothesise that blood pressure and obesity have causal roles in HCM. This is based on the findings of observational epidemiological studies that have noted a higher prevalence of hypertension and obesity in HCM individuals.[56, 340, 349] Establishing whether these heritable traits are involved in disease pathogenesis, or are an unintended sequelae of HCM, as individuals with HCM are recommended not to exercise intensely, can be addressed through two sample Mendelian randomisation.[351] Similarly, the relationship between HCM and other phenotypes of interest, such as heart failure or dilated cardiomyopathy can be evaluated. Furthermore, the observation that men were over represented in the HCM case-series stimulates further hypotheses regarding possible protective factors in women. There is evidence in the literature of sex dimorphism for complex traits, including body fat distribution.[356] Future analyses could further evaluate whether gender differences contribute towards the genetic architecture of HCM.

7.6.2 The role of common variants in the expressivity of disease

Second, to evaluate the contribution of common genetic variants towards the variable expressivity and penetrance of pathogenic variants in HCM. I would hypothesise that an individual's genetic risk score explains a large proportion of the phenotypic variability, and thus contributes towards understanding of variable penetrance and expressivity. Proving this may be challenging, given the high allelic heterogeneity in HCM, and analyses should focus on large groups of pathogenic variants that induce similar effects (i.e. truncating variants in *MYBCP3* or missense variants

in *MYH7*). Furthermore, accurate phenotypic assessments will be required to characterise the HCM phenotype, beyond the present binary classification used in these GWAS efforts.

7.6.3 Additional monogenic gene discovery

Third, in general, sarcomere negative individuals possess higher genetic risk scores than sarcomere positive individuals. It is plausible that a small proportion of sarcomere-negative individuals possess a rare variant in a yet to be defined disease gene. Therefore, I would hypothesise that future monogenic causes of HCM will be detected, in sarcomere negative individuals, in possession of a low HCM genetic risk score. Whilst many sarcomere negative individuals may possess a low HCM genetic risk score, individuals with a family history of HCM should be prioritised for family-based sequencing and subsequent analyses directed towards the identification of rare causal variants in a novel HCM gene.

7.6.4 From variant to function to therapy

Whilst GWAS has been successful in identifying common susceptibility variants, it has proven more challenging to appreciate how these variants influence disease-risk. Central to this is the observation that 90% of all genome-wide significant variants are located in the non-coding genome. [319] Therefore, of the genome-wide significant results derived from this thesis, the missense variant rs41306688 located in *ADPRHL1*, appears most tractable (given that it is a coding variant) for functional characterisation using *in vitro* or *in vivo* models. Non-coding variants are thought to regulate gene-expression, by modifying promoter or enhancer activity, or disrupting transcription factor binding, often with cell-type specific effects.[319] As such, mechanistic insights for non-coding genome-wide significant variants may require future analytical approaches that combine supporting data relating to genome accessibility, enhancer/promoter activity, and gene expression data.[319] Several statistical approaches have been developed to combine these various layers of information, including *fGWAS* as outlined by Pickrell (2014).[357]

When considering whether a disease-associated variant might be a tractable novel drug target, the functional characterisation of the variant is essential, but not sufficient. Understanding the mechanism underpinning the disease association is important, as it will determine which therapeutic modality can be leveraged. For instance, if a protein-truncating variant is shown to confer protection from disease a future therapeutic would seek to inhibit or degrade the protein or mRNA to recapitulate the desired protein-truncating effects.[281] However, alongside an understanding of the mechanism, information needs to be collected regarding the anticipated magnitude of perturbation that would be required to generate a clinically meaningful response, and how the effects of a therapeutic could be monitored (i.e. the identification of a biomarker).

7.7 Limitations

The results generated in this thesis, specifically relating to the common variant contribution towards HCM, provide a foundation upon which future studies addressing clinical outcomes can be performed. However, there are limitations with the current approach. For example, the case cohorts remain relatively small and this results in limited discovery power for sarcomere-positive/negative analyses. Future studies that incorporate larger case cohorts will be able to refine effect estimates and may detect additional disease-associated variants of interest. Such studies will be important, not only to validate the results outlined here, but to consider whether the care HCM patients, and their families, receive should be stratified based on the different biological mechanisms that are perturbed.

7.8 Conclusion

HCM is the most common, serious, genetic heart muscle disorder and a leading cause of sudden death. Genetic testing for rare, disease-causing, genetic variants in sarcomere genes is the standard of care and conducted at scale. However, ∼74% of HCM patients do not carry identifiable pathogenic or likely pathogenic variants

and, in those that do, there is substantial variation in penetrance and disease expression. Here, through the evaluation of monogenic, oligogenic and polygenic models of HCM, I have been able to provide sufficient evidence to reject the broad null hypothesis that the genetic architecture of HCM is restricted to known rare variant contributions. The overall findings presented here have the potential to impact clinical diagnostic genetic testing and influence understanding regarding the genetic aetiology of HCM, particularly with respect to common genetic variants.

Appendices

A

Contents

A.1 Cohorts

Descriptor	HCM case cohorts		HCM cases and controls		Control cohort	Replication cohorts			Reference materials	
Cohort	OMGL	HCMR	BRRD	UKBB	T2DM	RBH	GEL	AMC	gnomAD	TOPMED
HCM cases	2,757	2,636	213	Total: 502,543	0	411	476	999	Exomes: 125,748 Genomes: 15,708	Genomes: 62,784
Controls	0	0	5801	HCM: 326	12,297	1211	37,515	2,117		
Sequenced	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		
Sequencing platform	Gene panel Illumina MiSeq	Gene panel Illumina MiSeq	Illumina HiSeq WGS	Illumina NovaSeq 6000 (~50k)	Multiple, exomes	TruSight Cardio, gene panel	Illumina WGS	Gene panel		
Genotyping	No	Yes	NA	Yes	NA	Yes	NA	Yes	No	No
Genotyping array	No	Affymetrix Precision Research Medicine Array	NA	Affymetrix UK Biobank Axiom® array	NA	Illumina OmniExpress	NA	Illumina Infinium BeadChip, Illumina OmniExpress and Global Screening Array.	NA	NA
Analyses cohort has contributed towards										
Rare variant										
Descriptive	X	X	X							
Haplotype analysis		X								
Burden testing	X	X			X					
Oligogenicity										
Aggregate burden			X							
Multiple variants			X							
MYBPC3 conditional analysis	X	X	X							X
Penetrance										
Secondary finding prevalence			X	X					X	X
Positive likelihood framework	X	X								X
Recruitment by genotype clinical study			X							
GWAS										
Heritability		X	X	X						
GWAS		X	X	X						
Meta-analysis		X	X	X*						
Sarcomere positive		X	X	X*						
Sarcomere-negative		X		X*						
Synthetic association		X		X*						
Genetic risk score		X		X*						
Replication						X	X	X		

Table A.1: Summary of cohorts used throughout the thesis *Only controls included in HCMR vs. UKBB GWAS were included. The UKBB case-control GWAS was not included in studies beyond the initial heritability estimates and GWAS.

A.2 Genomic regions considered

Chr	Start	End	Gene	Ensembl gene ID	Ensembl transcript ID
1	156084699	156108908	*LMNA*	ENSG00000160789	ENST00000368300
1	201328327	201342393	*TNNT2*	ENSG00000118194	ENST00000367318
1	236849963	236925930	*ACTN2*	ENSG00000077522	ENST00000366578
2	105977729	106002984	*FHL2*	ENSG00000115641	ENST00000358129
2	179391728	179669380	*TTN*	ENSG00000155657	ENST00000589042
2	220283174	220290723	*DES*	ENSG00000175084	ENST00000373960
3	14166683	14183306	*TMEM43*	ENSG00000170876	ENST00000306077
3	38591801	38674809	*SCN5A*	ENSG00000183873	ENST00000413689
3	46899723	46904891	*MYL3*	ENSG00000160808	ENST00000395869
3	52485280	52488042	*TNNC1*	ENSG00000114854	ENST00000232975
6	7542138	7586122	*DSP*	ENSG00000096696	ENST00000379802
6	118869545	118880254	*PLN*	ENSG00000198523	ENST00000357525
7	128470681	128498588	*FLNC*	ENSG00000128591	ENST00000325888
7	151254276	151573716	*PRKAG2*	ENSG00000106617	ENST00000287878
10	75757955	75877938	*VCL*	ENSG00000035403	ENST00000211998
10	92672612	92680795	*ANKRD1*	ENSG00000148677	ENST00000371697
10	112404202	112595747	*RBM20*	ENSG00000203867	ENST00000369519
10	121411177	121436805	*BAG3*	ENSG00000151929	ENST00000369085
11	19204206	19214006	*CSRP3*	ENSG00000129170	ENST00000533783
11	47353411	47374209	*MYBPC3*	ENSG00000134571	ENST00000545968
11	111779477	111782459	*CRYAB*	ENSG00000109846	ENST00000533475
12	32945347	33049676	*PKP2*	ENSG00000057294	ENST00000340811
12	111348870	111358344	*MYL2*	ENSG00000111245	ENST00000228841
14	23882052	23902952	*MYH7*	ENSG00000092054	ENST00000355349
15	35082602	35087020	*ACTC1*	ENSG00000159251	ENST00000290378
15	63335018	63358109	*TPM1*	ENSG00000140416	ENST00000403994
17	39911985	39928117	*JUP*	ENSG00000173801	ENST00000565780
18	28647970	28681945	*DSC2*	ENSG00000134755	ENST00000280904
18	29078204	29126717	*DSG2*	ENSG00000046604	ENST00000261590
18	29171855	29178649	*TTR*	ENSG00000118271	ENST00000237014
19	55663191	55668968	*TNNI3*	ENSG00000129991	ENST00000344887
X	31140025	33358226	*DMD*	ENSG00000198947	ENST00000357033
X	100652786	100662902	*GLA*	ENSG00000102393	ENST00000218516
X	119562328	119603035	*LAMP2*	ENSG00000005893	ENST00000200639
X	135288555	135292195	*FHL1*	ENSG00000022267	ENST00000394155
X	153640170	153649354	*TAZ*	ENSG00000102125	ENST00000299328

Table A.2: Genomic regions evaluated through gene panel sequencing 35 genomic regions, mapped to GRCh37, selectively captured and amplified during OMGL/HCMR gene panel sequencing.

A.3 False positive variants identified during visual inspection of BAM files

Chr	Position	Ref	Alt
1	201331116	C	T
1	201331516	G	A
1	201333503	C	T
1	201337298	C	T
1	201337316	G	T
3	46904763	A	T
3	46904778	C	T
3	46904816	G	A
3	46904880	T	C
7	151267266	G	T
7	151292479	G	T
7	151478267	C	T
7	151478285	G	T
11	47354827	G	A
11	47354875	C	G
11	47354875	C	T
11	47354878	G	A
11	47354882	T	TG
11	47355229	G	T
11	47356589	T	C
11	47356592	C	T
11	47356593	G	A
11	47356632	CAG	C
11	47356649	G	A
11	47356667	A	G
11	47356670	C	T
11	47356671	G	A
11	47356685	GC	G
11	47356715	G	A
11	47356745	G	T
11	47356752	A	G
11	47358988	G	GA
11	47359308	G	A
11	47360873	A	T
11	47360898	C	T
11	47361269	A	C
11	47361298	G	T
11	47371368	C	A
11	47371570	G	A
11	47371619	CG	C
11	47371651	G	A
11	47371654	G	A
11	47372816	T	G
11	47372843	G	C
14	23884476	C	T
14	23884930	G	A
14	23886127	G	A
14	23886148	C	T
14	23886712	C	A
14	23886724	C	G
14	23886749	G	A
14	23886757	AT	A
14	23886761	G	A
14	23886777	C	T
14	23886808	G	C
14	23886810	G	A
14	23886870	C	T
14	23889398	C	A
14	23889399	GGT	G
14	23893358	C	T
14	23894175	G	A
14	23894176	C	A
14	23898184	C	A
14	23898200	TA	T
14	23898292	G	C
14	23900811	C	T
15	63336235	G	A
15	63336241	G	A
15	63336248	T	C
15	63336253	A	G
15	63353093	A	T
X	100658842	T	C
X	100662727	G	T
X	100662740	A	T
X	119562386	A	T
X	119562479	C	T
X	119565190	A	T
X	119565194	C	T
X	119575653	G	A
X	119581879	C	A
X	119582945	G	T

Table A.3: List of false positive variants False positive variants identified and excluded from analysis through manual evaluation of BAM files

A.4 Gene-level framework for parsing variants of uncertain significance

Gene (Class of variant established as pathogenic)	Effect	Literature based case vs. control analyses			Comment	VUS re-designation strategy
		All rare variants	Truncating variant	Non-truncating variants		
ACTC1 (Missense)	Excess in HCM	0.46%	No evidence supporting enrichment in cases.	0.46%	Rare variants in this gene are enriched in HCM cases. Truncating variants are extremely rare, both in HCM cases and controls. Currently no evidence to support pathogenicity of this class of variant in HCM. No data to support enrichment of missense variants in any region/domain of the protein.	There is a high prior probability that a rare variant detected in this gene in an individual with HCM is pathogenic. Therefore, assuming there is no conflicting evidence, novel rare variants in this gene have been assigned the classification VUS-favouring pathogenic.
	Odds Ratio [95% CI]	8 [5-14]		9 [5-15]		
	Etiological Fraction [95% CI]	0.88 [0.78-0.93]		0.88 [0.79-0.93]		
MYBPC3 (Predominantly truncating)	Excess in HCM	17%	9%	8%	Rare truncating variants strongly associated with HCM. Most frequent pathogenic variant is a missense variant (p.R502W). Other recurrent pathogenic missense variants reported.	High prior probability that a rare truncating variant in this gene is pathogenic (most assigned Pathogenic). Novel missense variants (MAF <0.004% in gnomAD), and no other supporting evidence, assigned VUS-indeterminate. Missense variant with a frequency >0.01% in Northern European population assigned VUS-favouring benign. Novel missense variants (MAF >0.01% in gnomAD) but enriched in HCM cases of same ethnicity (OR >30) assigned VUS-favouring pathogenic.
	Odds Ratio [95% CI]	12 [11-13]	118 [86-169]	6 [5-6]		
	Etiological Fraction [95% CI]	0.91 [0.91-0.92]	0.99 [0.99-0.99]	0.82 [0.80-0.84]		

MYH7 (Missense)	Excess in HCM	12.8%	No evidence supporting enrichment in HCM cases.	12.9%	Rare missense variants in this gene are enriched in HCM cases; empirical data to indicate that variants cluster in HCM cases in codons 181-937*(which encode the functionally important motor domain, lever arm and part of the rod).	Novel rare variant in codons 181-937 are assigned VUS-favouring pathogenic. If other conflicting evidence such as lack of conservation, minor amino acid change-assigned VUS-indeterminate. Novel rare variant outside 181-937 assigned VUS-indeterminate. Novel rare variant outside 181-937, lack of conservation, minor amino acid change assigned VUS-favouring benign. Variant in codon 181-937, not previously reported in cases (or reported in a small number of cases), no strong evidence of pathogenicity (previously reported, segregation, good biochemical evidence) and found in gnomAD Euro (NF) MAF >0.004% assigned VUS-favour benign.
	Odds Ratio [95% CI]	12 [11-13]		12 [11-13]		
	Etiological Fraction [95% CI]	0.91 [0.91-0.92] **EF(codons 181-937):0.97*		0.92 [0.91-0.92] **EF(codons 181-937):0.97*		
TNNT2 (Missense)	Excess in HCM	1.71%	0.15%	1.56%	Rare variants in this gene are enriched in HCM cases. No data (as yet) to support enrichment of missense variants in any specific region/domain of the protein. Currently no strong evidence to support pathogenicity of truncating variants in HCM. However, pathogenic nonsense variants in C-terminal region of the gene have been reported; these are expected to escape NMD and therefore not	Novel missense variant in this gene (not detected in controls), conserved amino acid, biochemically different amino acids=assigned VUS-favouring pathogenic. Missense variant in this gene, MAF <0.01% in controls, but enriched in cases (OR >10), conserved amino acid, biochemically different amino acids=assigned VUS-indeterminate. Missense variant, MAF < 0.01% in controls, no evidence of enrichment in cases and or not
	Odds Ratio [95% CI]	8 [6-10]	6[3-14]	8 [7-11]		
	Etiological Fraction [95% CI]	0.88 [0.84-0.91]	0.83 [0.61-0.93]	0.88 [0.84-0.91]		

					considered to cause haploinsufficiency. Pathogenic in-frame amino acid deletions have been found in individuals with DCM.	conserved/small amino acid change=VUS-favouring benign.
TNNI3 (Missense)	Excess in HCM	2%	No evidence supporting enrichment in HCM cases.	1.95%	Rare variants in this gene are enriched in HCM cases. Evidence of enrichment of missense variants in exons 7 and 8 of this gene (codons 125-210) in HCM cases compared to controls. Currently no strong evidence to support pathogenicity of truncating variants in HCM.	Novel missense variant in codon 125-210 (not detected in controls), conserved amino acid, biochemically different amino acids =assigned VUS-favouring pathogenic. Missense variant in codon 125-210, MAF <0.01% in controls, but enriched in case (OR >10), conserved amino acid, biochemically different amino acids=assigned VUS-indeterminate. Missense variant not in codon 125-210, MAF < 0.01% in controls, but no evidence of case enrichment and or not conserved/small amino acid change=VUS-favouring benign.
	Odds Ratio [95% CI]	10 [8-13]		12 [9-15]		
	Etiological Fraction [95% CI]	0.90 [0.87-0.92]		0.91 [0.89-0.93]		
MYL3 (Missense)	Excess in HCM	0.7%	No evidence supporting enrichment in HCM cases.	0.71%	Rare missense variants in this gene are enriched in HCM cases. No data to support enrichment of missense variants in any specific region/domain of the protein.	Caution with novel rare variants in this gene. No reliable criteria which can be used to assign 'VUS-favouring pathogenic' classification.
	Odds Ratio [95% CI]	5 [3-7]		5 [3-7]		
	Etiological Fraction [95% CI]	0.80 [0.7-0.86]		0.8 [0.7-0.86]		
MYL2 (Missense)	Excess in HCM	0.92%	No evidence supporting enrichment in HCM cases.	0.87%	Rare missense variants in this gene are enriched in HCM cases. No data to support enrichment of missense variants in any specific region/domain of the protein.	Caution with novel rare variants in this gene. No reliable criteria which can be used to assign 'VUS-favouring pathogenic' classification.
	Odds Ratio [95% CI]	6 [4-9]		7 [5-10]		
	Etiological Fraction [95% CI]	0.84 [0.77-0.89]		0.85 [0.78-0.90]		

TPM1 (Missense)	Excess in HCM	1.4%	No evidence supporting enrichment in HCM cases.	1.4%	Rare variants in this gene are enriched in HCM cases. Truncating variants are extremely rare, both in HCM cases and controls. Currently no evidence to support pathogenicity of this class of variant in HCM. No data to support enrichment of missense variants in any region/domain of the protein.	There is a high prior probability that a rare variant detected in this gene in an individual with HCM is pathogenic. Therefore, assuming there is no conflicting evidence, novel rare variants in this gene have been assigned the classification VUS-favouring pathogenic.
	Odds Ratio [95% CI]	17 [12-25]		18 [12-26]		
	Etiological Fraction [95% CI]	0.94 [0.92-0.96]		0.94 [0.92-0.96]		

Table A.4: Summary of gene-based approach taken to re-classify variants of uncertain significance so that HCM cases could be dichotomised into sarcomere-positive and sarcomere-negative groups.

A.5 Penetrance estimates

Gene	HGVS.c	HGVS.p	OR [95% CI]	Post-test probability [95% CI]	ACMG class
MYBPC3	c.1504C>T	p.Arg502Trp	277 (154 - 491)	0.36 (0.24-0.5)	5
MYBPC3	c.772G>A	p.Glu258Lys	304 (142 - 645)	0.38 (0.22-0.56)	5
MYBPC3	c.1224-52G>A	-	780 (107 - 5600)	0.61 (0.18-0.92)	5
MYBPC3	c.1624+4A>T	-	633 (139 - 2860)	0.56 (0.22-0.85)	5
MYBPC3	c.1624G>C	p.Glu542Gln	97.5 (47.3 - 200)	0.16 (0.09-0.29)	5
MYH7	c.2389G>A	p.Ala797Thr	136 (59.5 - 309)	0.21 (0.11-0.38)	5
MYBPC3	c.1484G>A	p.Arg495Gln	136 (56.5 - 324)	0.21 (0.1-0.39)	4
MYH7	c.1988G>A	p.Arg663His	180 (58.6 - 552)	0.27 (0.11-0.53)	5
MYH7	c.5135G>A	p.Arg1712Gln	101 (39.1 - 260)	0.17 (0.07-0.34)	4
MYBPC3	c.3330+5G>C	-	30.6 (15.3 - 61.3)	0.06 (0.03-0.11)	5
MYBPC3	c.927-9G>A	-	151 (44.2 - 514)	0.23 (0.08-0.51)	5
TNNI3	c.433C>T	p.Arg145Trp	172 (49.2 - 600)	0.26 (0.09-0.55)	5
MYBPC3	c.3330+2T>G	-	300 (47.8 - 1870)	0.38 (0.09-0.79)	5
MYBPC3	c.2864_2865del	p.Pro955ArgfsTer95	426 (59.9 - 3030)	0.46 (0.11-0.86)	5
MYBPC3	c.1227-13G>A	-	114 (37.5 - 347)	0.19 (0.07-0.41)	5
MYH7	c.2681A>G	p.Glu894Gly	217 (19.5 - 2400)	0.3 (0.04-0.83)	5
MYH7	c.2609G>A	p.Arg870His	157 (45.1 - 546)	0.24 (0.08-0.52)	5
MYBPC3	c.2308G>A	p.Asp770Asn	77.9 (23.6 - 257)	0.13 (0.05-0.34)	4
MYBPC3	c.1224-19G>A	-	41.8 (14.1 - 124)	0.08 (0.03-0.2)	4
MYH7	c.2722C>G	p.Leu908Val	82.1 (16.3 - 413)	0.14 (0.03-0.45)	5
TNNI3	c.422G>A	p.Arg141Gln	24.6 (8.81 - 68.8)	0.05 (0.02-0.12)	4
MYBPC3	c.3226_3227insT	p.Asp1076ValfsTer6	170 (16 - 1800)	0.25 (0.03-0.78)	5
MYL2	c.173G>A	p.Arg58Gln	128 (23.8 - 690)	0.2 (0.05-0.58)	5
MYH7	c.4130C>T	p.Thr1377Met	860 (79.8 - 9230)	0.63 (0.14-0.95)	5
MYH7	c.2606G>A	p.Arg869His	46 (19.8 - 106)	0.08 (0.04-0.18)	4
TNNI3	c.485G>A	p.Arg162Gln	26.9 (11.2 - 64.5)	0.05 (0.02-0.11)	4
MYBPC3	c.2827C>T	p.Arg943Ter	79.2 (22.4 - 279)	0.14 (0.04-0.36)	5
MYBPC3	c.1483C>G	p.Arg495Gly	188 (32.1 - 1100)	0.27 (0.06-0.69)	4
MYL2	c.64G>A	p.Glu22Lys	39.1 (12.5 - 123)	0.07 (0.02-0.2)	5
MYH7	c.2717A>G	p.Asp906Gly	1230 (30.1 - 50000)	0.71 (0.06-0.99)	5
MYBPC3	c.1505G>A	p.Arg502Gln	44.3 (8.73 - 225)	0.08 (0.02-0.31)	5
MYBPC3	c.927-2A>G	-	82.7 (15.6 - 438)	0.14 (0.03-0.47)	5
MYBPC3	c.821+1G>A	-	75.2 (24.2 - 233)	0.13 (0.05-0.32)	5
MYH7	c.4066G>A	p.Glu1356Lys	97.2 (10.1 - 936)	0.16 (0.02-0.65)	5
MYH7	c.1816G>A	p.Val606Met	104 (24.4 - 440)	0.17 (0.05-0.47)	5
TNNI3	c.484C>T	p.Arg162Trp	20.7 (8.09 - 52.9)	0.04 (0.02-0.1)	4
MYBPC3	c.3190+5G>A	-	34.5 (9.48 - 125)	0.07 (0.02-0.2)	4
MYH7	c.4135G>A	p.Ala1379Thr	216 (18.1 - 2570)	0.3 (0.04-0.84)	5
MYH7	c.2167C>T	p.Arg723Cys	49.8 (11.9 - 208)	0.09 (0.02-0.29)	5
MYH7	c.428G>A	p.Arg143Gln	187 (16.4 - 2120)	0.27 (0.03-0.81)	4
MYH7	c.427C>T	p.Arg143Trp	20.2 (6.72 - 60.7)	0.04 (0.01-0.11)	4
TNNI3	c.586G>A	p.Asp196Asn	141 (24.2 - 817)	0.22 (0.05-0.62)	4
TNNT2	c.266T>A	p.Ile89Asn	103 (17.9 - 598)	0.17 (0.04-0.55)	5
MYBPC3	c.3697C>T	p.Gln1233Ter	143 (30.2 - 676)	0.22 (0.06-0.58)	5
MYBPC3	c.3065G>C	p.Arg1022Pro	23.8 (7.7 - 73.4)	0.05 (0.02-0.13)	4
MYBPC3	c.2459G>A	p.Arg820Gln	45.5 (14.6 - 142)	0.08 (0.03-0.22)	4
MYBPC3	c.2374T>C	p.Trp792Arg	81.1 (20.5 - 320)	0.14 (0.04-0.39)	4
MYH7	c.1750G>C	p.Gly584Arg	147 (13.4 - 1600)	0.23 (0.03-0.76)	5
MYL3	c.427G>A	p.Glu143Lys	39.7 (11.2 - 140)	0.07 (0.02-0.22)	4
TNNT2	c.890G>A	p.Trp297Ter	77.3 (13 - 459)	0.13 (0.03-0.48)	5
MYBPC3	c.3811C>T	p.Arg1271Ter	36.3 (8.85 - 149)	0.07 (0.02-0.23)	4
MYBPC3	c.3233G>A	p.Trp1078Ter	92.3 (15.3 - 557)	0.16 (0.03-0.53)	5
MYBPC3	c.1790G>A	p.Arg597Gln	31.9 (9.19 - 110)	0.06 (0.02-0.18)	4
MYH7	c.2605C>T	p.Arg869Cys	24.4 (4.35 - 137)	0.05 (0.01-0.21)	4
TNNI3	c.497C>T	p.Ser166Phe	114 (23.3 - 560)	0.19 (0.05-0.53)	4
TNNI3	c.434G>A	p.Arg145Gln	23.8 (6.63 - 85.3)	0.05 (0.01-0.15)	4
TNNT2	c.304C>T	p.Arg102Trp	61.9 (9.7 - 395)	0.11 (0.02-0.44)	5
MYBPC3	c.3776del	p.Gln1259ArgfsTer72	79.8 (12.3 - 519)	0.14 (0.02-0.51)	5
MYBPC3	c.3613C>T	p.Arg1205Trp	56.8 (8.95 - 361)	0.1 (0.02-0.42)	4
MYBPC3	c.3190+2T>G	-	72.3 (11.2 - 466)	0.13 (0.02-0.48)	5
MYBPC3	c.2610del	p.Ser871AlafsTer8	8.22 (2.29 - 29.5)	0.02 (0.01-0.06)	5
MYBPC3	c.2458C>T	p.Arg820Trp	179 (14 - 2290)	0.26 (0.03-0.82)	4
MYBPC3	c.1591G>A	p.Gly531Arg	14.3 (4.12 - 49.4)	0.03 (0.01-0.09)	4

MYBPC3	c.436dup	p.Thr146AsnfsTer7	105 (9.18 - 1190)	0.17 (0.02-0.7)	5
MYH7	c.2348G>A	p.Arg783His	25.8 (6.39 - 105)	0.05 (0.01-0.17)	4
MYH7	c.2221G>A	p.Gly741Arg	46.2 (4.45 - 478)	0.09 (0.01-0.49)	5
MYH7	c.2155C>T	p.Arg719Trp	46.2 (4.45 - 478)	0.09 (0.01-0.49)	5
MYH7	c.715G>A	p.Asp239Asn	78.7 (7.2 - 860)	0.14 (0.01-0.63)	4
MYH7	c.610C>T	p.Arg204Cys	40.7 (9.98 - 166)	0.08 (0.02-0.25)	4
ACTC1	c.301G>A	p.Glu101Lys	78.7 (7.2 - 860)	0.14 (0.01-0.63)	5
TPM1	c.523G>A	p.Asp175Asn	16.3 (4.18 - 63.3)	0.03 (0.01-0.11)	5
MYL3	c.517A>G	p.Met173Val	77.4 (14.7 - 407)	0.13 (0.03-0.45)	4
TNNT2	c.418C>T	p.Arg140Cys	182 (12.6 - 2640)	0.27 (0.03-0.84)	4
MYBPC3	c.3624del	p.Lys1209SerfsTer28	53.5 (7.16 - 400)	0.1 (0.01-0.44)	5
MYBPC3	c.2555dup	p.Gly853ArgfsTer31	30.7 (5.02 - 187)	0.06 (0.01-0.27)	5
MYBPC3	c.2258dup	p.Lys754GlufsTer79	29.5 (5.65 - 154)	0.06 (0.01-0.24)	5
MYBPC3	c.1800del	p.Lys600AsnfsTer2	89 (7.03 - 1120)	0.15 (0.01-0.69)	5
MYBPC3	c.1227-2A>G	-	29.8 (2.58 - 345)	0.06 (0.01-0.41)	5
MYBPC3	c.833del	p.Gly278GlufsTer22	42.2 (6.85 - 260)	0.08 (0.01-0.34)	5
MYBPC3	c.292+1G>A	-	30.8 (2.67 - 356)	0.06 (0.01-0.42)	5
MYH7	c.3158G>A	p.Arg1053Gln	4.12 (0.954 - 17.7)	0.01 (0-0.03)	4
MYH7	c.2765T>C	p.Met922Thr	29.7 (5.69 - 155)	0.06 (0.01-0.24)	4
MYH7	c.2080C>T	p.Arg694Cys	18.2 (3.84 - 86)	0.04 (0.01-0.15)	4
MYH7	c.1324C>T	p.Arg442Cys	9.36 (1.98 - 44.2)	0.02 (0-0.08)	4
MYH7	c.1231G>A	p.Val411Ile	8.52 (1.71 - 42.3)	0.02 (0-0.08)	4
MYH7	c.343T>C	p.Tyr115His	108 (8.28 - 1400)	0.18 (0.02-0.74)	4
MYBPC3	c.3642G>A	p.Trp1214Ter	25 (1.51 - 412)	0.05 (0-0.45)	4
MYBPC3	c.3286G>T	p.Glu1096Ter	39.5 (2.34 - 665)	0.07 (0-0.57)	5
MYBPC3	c.3181C>T	p.Gln1061Ter	12.1 (1.34 - 108)	0.02 (0-0.18)	5
MYBPC3	c.2490dup	p.His831SerfsTer2	18.1 (2 - 163)	0.04 (0-0.25)	5
MYBPC3	c.2308+1G>A	-	25 (1.51 - 412)	0.05 (0-0.45)	5
MYBPC3	c.2113dup	p.Thr705AsnfsTer3	50.6 (2.96 - 866)	0.09 (0.01-0.63)	5
MYBPC3	c.1357_1358del	p.Pro453CysfsTer21	22.5 (2 - 252)	0.04 (0-0.34)	5
MYBPC3	c.1223+1G>A	-	12.3 (1.27 - 119)	0.02 (0-0.19)	5
MYBPC3	c.1090+1G>T	-	50.5 (2.95 - 864)	0.09 (0.01-0.63)	5
MYBPC3	c.1038_1042dup	p.Met348ThrfsTer4	50.5 (2.95 - 865)	0.09 (0.01-0.63)	5
MYBPC3	c.787G>A	p.Gly263Arg	2.18 (0.297 - 16)	0 (0-0.03)	4
MYBPC3	c.747C>A	p.Cys249Ter	29.2 (1.75 - 484)	0.06 (0-0.49)	5
MYBPC3	c.743_746del	p.Asp248AlafsTer51	47.7 (2.8 - 813)	0.09 (0.01-0.62)	5
MYBPC3	c.624G>C	p.Gln208His	0.936 (0.13 - 6.73)	0 (0-0.01)	4
MYH7	c.5458C>T	p.Arg1820Trp	10.5 (1.29 - 85.4)	0.02 (0-0.15)	4
MYH7	c.2608C>T	p.Arg870Cys	9.02 (1.05 - 77.5)	0.02 (0-0.13)	4
MYH7	c.2555T>C	p.Met852Thr	30.6 (1.84 - 509)	0.06 (0-0.5)	4
MYH7	c.2378G>A	p.Arg793Gln	23.4 (2.4 - 228)	0.05 (0-0.31)	4
MYH7	c.2248G>C	p.Asp750His	48.6 (2.85 - 829)	0.09 (0.01-0.62)	4
MYH7	c.2200C>G	p.Gln734Glu	51 (2.98 - 874)	0.09 (0.01-0.64)	4
MYH7	c.1954A>G	p.Arg652Gly	51.1 (2.98 - 875)	0.09 (0.01-0.64)	5
MYH7	c.1805A>G	p.Asn602Ser	51.1 (2.98 - 875)	0.09 (0.01-0.64)	4
MYH7	c.1727A>G	p.His576Arg	12.9 (1.54 - 107)	0.03 (0-0.18)	4
MYH7	c.1491G>T	p.Glu497Asp	18.2 (2.02 - 164)	0.04 (0-0.25)	4
MYH7	c.1447G>A	p.Glu483Lys	38.7 (3.41 - 440)	0.07 (0.01-0.47)	5
MYH7	c.949G>C	p.Glu317Gln	48.6 (2.85 - 830)	0.09 (0.01-0.62)	4
MYH7	c.788T>C	p.Ile263Thr	48.6 (2.85 - 830)	0.09 (0.01-0.62)	5
MYH7	c.727C>T	p.Arg243Cys	29.2 (1.75 - 485)	0.06 (0-0.49)	4
MYH7	c.532G>A	p.Gly178Arg	23.2 (2.07 - 260)	0.04 (0-0.34)	4
TNNI3	c.596G>A	p.Ser199Asn	25 (1.51 - 412)	0.05 (0-0.45)	4
TNNI3	c.592C>G	p.Leu198Val	50.7 (2.96 - 868)	0.09 (0.01-0.63)	4
TNNI3	c.556C>T	p.Arg186Trp	12.2 (1.46 - 102)	0.02 (0-0.17)	4
TNNI3	c.470C>T	p.Ala157Val	12.2 (1.09 - 135)	0.02 (0-0.21)	4
TNNI3	c.407G>A	p.Arg136Gln	50.5 (2.95 - 864)	0.09 (0.01-0.63)	4
MYL3	c.281G>A	p.Arg94His	23.2 (2.06 - 260)	0.04 (0-0.34)	4

Table A.5: Extended list of penetrance estimates for secondary findings Variants identified as pathogenic or likely pathogenic variants in 8 core sarcomere genes, specific to HCM, were evaluated.

References

[1] Andrew R. Harper, Arash Yavari, and Houman Ashrafian. “Inherited cardiomyopathies”. In: *Medicine (United Kingdom)* (2014).

[2] Hugh Watkins, Houman Ashrafian, and Charles Redwood. “Inherited cardiomyopathies”. In: *New England Journal of Medicine* 364.17 (Apr. 2011). Ed. by Robert S. Schwartz, pp. 1643–1656. URL: http://www.nejm.org/doi/10.1056/NEJMra0902923.

[3] Barry J. Maron et al. “Prevalence of hypertrophic cardiomyopathy in a general population of young adults: Echocardiographic analysis of 4111 subjects in the CARDIA study”. In: *Circulation* 92.4 (Aug. 1995), pp. 785–789.

[4] Yubao Zou et al. “Prevalence of idiopathic hypertrophic cardiomyopathy in China: A population-based echocardiographic analysis of 8080 adults”. In: *American Journal of Medicine* 116.1 (Jan. 2004), pp. 14–18. URL: http://www.ncbi.nlm.nih.gov/pubmed/14706660.

[5] Hiroyuki Morita et al. “Single-gene mutations and increased left ventricular wall thickness in the community: the Framingham Heart Study.” In: *Circulation* 113.23 (June 2006), pp. 2697–705. URL: http://www.ncbi.nlm.nih.gov/pubmed/16754800.

[6] Hee Hwa Ho et al. “Clinical characteristics of and long-term outcome in chinese patients with hypertrophic cardiomyopathy”. In: *American Journal of Medicine* (2004).

[7] Barry J. Maron, Ethan J. Rowin, and Martin S. Maron. “Global Burden of Hypertrophic Cardiomyopathy”. In: *JACC: Heart Failure* (2018).

[8] Roddy Walsh et al. “Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples.” ENG. In: *Genetics in medicine : official journal of the American College of Medical Genetics* (Aug. 2016).

[9] Ali J. Marian. “Sudden cardiac death in patients with hypertrophic cardiomyopathy: From bench to bedside with an emphasis on genetic markers”. In: *Clinical Cardiology* 18.4 (Apr. 1995), pp. 189–198. URL: http://doi.wiley.com/10.1002/clc.4960180403.

[10] Barry J. Maron et al. “Hypertrophic cardiomyopathy in adulthood associated with low cardiovascular mortality with contemporary management strategies”. In: *Journal of the American College of Cardiology* (2015).

[11] Richard D. Bagnall et al. “A Prospective Study of Sudden Cardiac Death among Children and Young Adults”. In: *New England Journal of Medicine* 374.25 (June 2016), pp. 2441–2452. URL: http://www.nejm.org/doi/10.1056/NEJMoa1510687.

[12] Jose Luis Zamorano et al. "2014 ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy: The task force for the diagnosis and management of hypertrophic cardiomyopathy of the European Society of Cardiology (ESC)". In: *European Heart Journal* 35.39 (Oct. 2014), pp. 2733–2779.

[13] Mar Pujades-Rodriguez et al. "Identifying unmet clinical need in hypertrophic cardiomyopathy using national electronic health records". In: *PLoS ONE* (2018).

[14] A.R. Harper et al. "Delivering Clinical Grade Sequencing and Genetic Test Interpretation for Cardiovascular Medicine". In: *Circulation: Cardiovascular Genetics* 10.2 (2017).

[15] Sarah Wordsworth et al. "DNA testing for hypertrophic cardiomyopathy: A cost-effectiveness model". In: *European Heart Journal* (2010).

[16] Andrew R. Harper, Hitesh C. Patel, and Alexander R. Lyon. "Heart failure with preserved ejection fraction". In: *Clinical Medicine, Journal of the Royal College of Physicians of London* 18.Suppl 2 (Apr. 2018), s24–ss29. URL: `https://www.rcpjournals.org/content/clinmedicine/18/Suppl_2/s24%20https://www.rcpjournals.org/content/clinmedicine/18/Suppl_2/s24.abstract`.

[17] Virginia S. Hahn et al. "Endomyocardial Biopsy Characterization of Heart Failure With Preserved Ejection Fraction and Prevalence of Cardiac Amyloidosis". In: *JACC: Heart Failure* (July 2020). URL: `https://linkinghub.elsevier.com/retrieve/pii/S2213177920302535`.

[18] D. TEARE. "Asymmetrical hypertrophy of the heart in young adults." In: *British heart journal* 20.1 (Jan. 1958), pp. 1–8.

[19] A Vulpian. "Contribution à l'étude des rétrécissements de l'orifice ventriculo-aortique". In: *Arch Physiol* 3 (1868), pp. 220–222.

[20] LIOUVILLE and H. "Retrecissement cardiaque sous mortique". In: *Gaz Med Paris* 24 (1869), pp. 161–163.

[21] HALLOPEAU and M. "Retrecissement ventriculoaortique". In: *Gaz Med Paris* 24 (1869), pp. 683–684.

[22] Amanda C. Garfinkel, Jonathan G. Seidman, and Christine E. Seidman. "Genetic Pathogenesis of Hypertrophic and Dilated Cardiomyopathy". In: *Heart Failure Clinics* 14.2 (Apr. 2018), pp. 139–146.

[23] W. J. McKenna and J. E. Deanfield. "Hypertrophic cardiomyopathy: An important cause of sudden death". In: *Archives of Disease in Childhood* (1984).

[24] Barry J. Maron et al. "How hypertrophic cardiomyopathy became a contemporary treatable genetic disease with low mortality: Shaped by 50 years of clinical research and practice". In: *JAMA Cardiology* (2016).

[25] Barry J. Maron et al. "Clinical course of hypertrophic cardiomyopathy with survival to advanced age". In: *Journal of the American College of Cardiology* (2003).

[26] Aurore Lyon et al. "Distinct ECG phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers". In: *Frontiers in Physiology* 9.MAR (Mar. 2018).

[27] Aurore Lyon et al. “Electrocardiogram phenotypes in hypertrophic cardiomyopathy caused by distinct mechanisms: Apico-basal repolarization gradients vs. Purkinje-myocardial coupling abnormalities”. In: *Europace* (2018).

[28] Wei Yin Ko et al. “Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram”. In: *Journal of the American College of Cardiology* 75.7 (Feb. 2020), pp. 722–733.

[29] Bernard J. Gersh et al. “2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: A report of the American College of cardiology foundation/American heart association task force on practice guidelines”. In: *Circulation* 124.24 (Dec. 2011).

[30] Rafaela Soler et al. “Phenotypes of hypertrophic cardiomyopathy. An illustrative review of MRI findings”. In: *Insights into Imaging* (2018).

[31] Rina Ariga et al. “Identification of Myocardial Disarray in Patients With Hypertrophic Cardiomyopathy and Ventricular Arrhythmias”. In: *Journal of the American College of Cardiology* 73.20 (May 2019), pp. 2493–2502.

[32] Zhen Weng et al. “Prognostic Value of LGE-CMR in HCM: A Meta-Analysis”. In: *JACC: Cardiovascular Imaging* (2016).

[33] Carolyn Y. Ho et al. “T1 measurements identify extracellular volume expansion in hypertrophic cardiomyopathy sarcomere mutation carriers with and without left ventricular hypertrophy”. In: *Circulation: Cardiovascular Imaging* (2013).

[34] Michael Salerno et al. “Recent Advances in Cardiovascular Magnetic Resonance”. In: *Circulation: Cardiovascular Imaging* (2017).

[35] John A. Jarcho et al. “Mapping a Gene for Familial Hypertrophic Cardiomyopathy to Chromosome 14q1”. In: *New England Journal of Medicine* (1989).

[36] Scott D. Solomon et al. “Familial hypertrophic cardiomyopathy is a genetically heterogeneous disease”. In: *Journal of Clinical Investigation* (1990).

[37] Ludwig Thierfelder et al. “α-tropomyosin and cardiac troponin T mutations cause familial hypertrophic cardiomyopathy: A disease of the sarcomere”. In: *Cell* (1994).

[38] A. Kimura et al. “Mutations in the cardiac troponin I gene associated with hypertrophic cardiomyopathy”. In: *Nature Genetics* (1997).

[39] Jens Mogensen et al. “α-cardiac actin is a novel disease gene in familial hypertrophic cardiomyopathy”. In: *Journal of Clinical Investigation* (1999).

[40] Karl Poetter et al. “Mutations in either the essential or regulatory light chains of myosin are associated with a rare myopathy in human heart and skeletal muscle”. In: *Nature Genetics* (1996).

[41] Hugh Watkins et al. “Mutations in the Genes for Cardiac Troponin T and α-Tropomyosin in Hypertrophic Cardiomyopathy”. In: *New England Journal of Medicine* (1995).

[42] Hugh Huxley and Jean Hanson. “Changes in the Cross-striations of muscle during contraction and stretch and their structural interpretation”. In: *Nature* (1954).

[43] A. F. Huxley and R. Niedergerke. “Structural changes in muscle during contraction: Interference microscopy of living muscle fibres”. In: *Nature* (1954).

[44] Regina El Dib et al. “Enzyme replacement therapy for Anderson-Fabry disease: A complementary overview of a Cochrane publication through a linear regression and a pooled analysis of proportions from cohort studies”. In: *PLOS ONE* 12.3 (Mar. 2017). Ed. by Leighton R James, e0173358. URL: https://dx.plos.org/10.1371/journal.pone.0173358.

[45] E. Blair. “Mutations in the gamma2 subunit of AMP-activated protein kinase cause familial hypertrophic cardiomyopathy: evidence for the central role of energy compromise in disease pathogenesis”. In: *Human Molecular Genetics* 10.11 (May 2001), pp. 1215–1220.

[46] Matthew G.D. Bates et al. “Cardiac involvement in mitochondrial DNA disease: Clinical spectrum, diagnosis, and management”. In: *European Heart Journal* (2012).

[47] Barry J. Maron et al. “Clinical outcome and phenotypic expression in LAMP2 cardiomyopathy”. In: *JAMA - Journal of the American Medical Association* 301.12 (Mar. 2009), pp. 1253–1259.

[48] Frederick L. Ruberg and John L. Berk. “Transthyretin (TTR) Cardiac Amyloidosis”. In: *Circulation* 126.10 (Sept. 2012), pp. 1286–1300. URL: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.111.078915.

[49] Christopher N. Toepfer et al. “Hypertrophic cardiomyopathy mutations in MYBPC3 dysregulate myosin”. In: *Science Translational Medicine* 11.476 (Jan. 2019).

[50] Christopher N. Toepfer et al. “Myosin Sequestration Regulates Sarcomere Function, Cardiomyocyte Energetics, and Metabolism, Informing the Pathogenesis of Hypertrophic Cardiomyopathy”. In: *Circulation* (2020).

[51] Roddy Walsh et al. “Defining the genetic architecture of hypertrophic cardiomyopathy: Re-evaluating the role of non-sarcomeric genes”. In: *European Heart Journal* (2017).

[52] Konrad J. Karczewski et al. “Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes”. In: *bioRxiv* (2019).

[53] Monkol Lek et al. “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* (2016).

[54] Daniel Taliun et al. “Sequencing of 53 , 831 diverse genomes from the NHLBI TOPMed Program”. In: *Biorxiv* (2019), pp. 1–46.

[55] Ahmed A. Alfares et al. “Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: Expanded panels offer limited additional sensitivity”. In: *Genetics in Medicine* (2015).

[56] Carolyn Y. Ho et al. “Genotype and lifetime burden of disease in hypertrophic cardiomyopathy”. In: *Circulation* (2018).

[57] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467.

[58] *Hypertrophic cardiomyopathy - teen and adult (Version 2.1)*. URL: https://panelapp.genomicsengland.co.uk/panels/49/.

[59] Jodie Ingles et al. "Evaluating the Clinical Validity of Hypertrophic Cardiomyopathy Genes". In: *Circulation: Genomic and Precision Medicine* 12.2 (Feb. 2019), pp. 57–64. URL: https://www.ahajournals.org/doi/10.1161/CIRCGEN.119.002460.

[60] Christian Geier et al. "Beyond the sarcomere: CSRP3 mutations cause hypertrophic cardiomyopathy". In: *Human Molecular Genetics* (2008).

[61] Francesca Girolami et al. "Novel α-Actinin 2 Variant Associated With Familial Hypertrophic Cardiomyopathy and Juvenile Atrial Arrhythmias". In: *Circulation: Cardiovascular Genetics* 7.6 (Dec. 2014), pp. 741–750. URL: https://www.ahajournals.org/doi/10.1161/CIRCGENETICS.113.000486.

[62] Rafael Valdés-Mas et al. "Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy". In: *Nature Communications* (2014).

[63] Rowida Almomani et al. "Biallelic Truncating Mutations in ALPK3 Cause Severe Pediatric Cardiomyopathy". In: *Journal of the American College of Cardiology* 67.5 (Feb. 2016), pp. 515–525. URL: https://linkinghub.elsevier.com/retrieve/pii/S0735109715076287.

[64] Sari U.M. Vanninen et al. "Heterozygous junctophilin-2 (JPH2) p. (Thr161Lys) is a monogenic cause for HCM with heart failure". In: *PLoS ONE* 13.9 (Sept. 2018), e0203422. URL: http://www.ncbi.nlm.nih.gov/pubmed/30235249%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6147424.

[65] Juan Pablo Ochoa et al. "Formin Homology 2 Domain Containing 3 (FHOD3) Is a Genetic Basis for Hypertrophic Cardiomyopathy". In: *Journal of the American College of Cardiology* 72.20 (Nov. 2018), pp. 2457–2467.

[66] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of age: Ten years of next-generation sequencing technologies". In: *Nature Reviews Genetics* 17.6 (June 2016), pp. 333–351. URL: www.nature.com/nrg.

[67] Jay Shendure et al. "DNA sequencing at 40: Past, present and future". In: *Nature* 550.7676 (Oct. 2017), pp. 345–353. URL: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

[68] Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. "High-Throughput Sequencing Technologies". In: *Molecular Cell* 58.4 (May 2015), pp. 586–597. URL: http://dx.doi.org/10.1016/j.molcel.2015.05.004.

[69] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing". In: *Nature Biotechnology* 26.10 (Oct. 2008), pp. 1135–1145. URL: http://www.nature.com/naturebiotechnology.

[70] Rachel L Goldfeder et al. "Practice of Epidemiology Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis". In: (2017). URL: https://academic.oup.com/aje/article-abstract/186/8/1000/3811717.

[71] Antonio Rueda Martin et al. "PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels". In: *Nature Genetics* 51.11 (Nov. 2019), pp. 1560–1565.

[72] Lili Li et al. "A Potential Oligogenic Etiology of Hypertrophic Cardiomyopathy: A Classic Single-Gene Disorder." In: *Circulation research* 120.7 (Mar. 2017), pp. 1084–1090. URL: http://www.ncbi.nlm.nih.gov/pubmed/28223422%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5380229.

[73] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. "The promise of whole-exome sequencing in medical genetics". In: *Journal of Human Genetics* 59.1 (Jan. 2014), pp. 5–15.

[74] Richard D. Bagnall et al. "Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients With Hypertrophic Cardiomyopathy". In: *Journal of the American College of Cardiology* 72.4 (July 2018), pp. 419–429.

[75] Cristina Barbosa, Isabel Peixeiro, and Luísa Romão. "Gene Expression Regulation by Upstream Open Reading Frames and Human Disease". In: *PLoS Genetics* 9.8 (Aug. 2013), e1003529. URL: http://www.ncbi.nlm.nih.gov/pubmed/23950723%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3738444.

[76] Nicola Whiffin et al. "Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals Aggregation Database (gnomAD) Production Team, Genome Aggregation Database (gnomAD) Consortium, Stuart A Cook". In: *bioRxiv* (Aug. 2019), p. 543504. URL: http://dx.doi.org/10.1101/543504.

[77] Patrick J. Short et al. "De novo mutations in regulatory elements in neurodevelopmental disorders". In: *Nature* 555.7698 (Mar. 2018), pp. 611–616. URL: http://www.ncbi.nlm.nih.gov/pubmed/29562236%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5912909.

[78] Euan A. Ashley. "Towards precision medicine". In: *Nature Reviews Genetics* 17.9 (Sept. 2016), pp. 507–522.

[79] Shanika L. Amarasinghe et al. "Opportunities and challenges in long-read sequencing data analysis". In: *Genome Biology* 21.1 (Feb. 2020), pp. 1–16.

[80] Doruk Beyter et al. "Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease". In: *bioRxiv* (Nov. 2019), p. 848366.

[81] Jason D. Merker et al. "Long-read genome sequencing identifies causal structural variation in a Mendelian disease". In: *Genetics in Medicine* 20.1 (Jan. 2018), pp. 159–163.

[82] Satomi Mitsuhashi and Naomichi Matsumoto. "Long-read sequencing for rare human genetic diseases". In: *Journal of Human Genetics* 65.1 (Jan. 2020), pp. 11–19.

[83] Anath C. Lionel et al. "Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test". In: *Genetics in Medicine* 20.4 (Apr. 2018), pp. 435–443. URL: http://www.ncbi.nlm.nih.gov/pubmed/28771251%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5895460.

[84] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* (2009).

[85] "E pluribus unum". In: *Nature Methods* 7.5 (May 2010), p. 331.

[86] Justin M. Zook et al. "Extensive sequencing of seven human genomes to characterize benchmark reference materials". In: *Scientific Data* (2016).

[87] Justin M. Zook et al. "An open resource for accurately benchmarking small variant and reference calls". In: *Nature Biotechnology* 37.5 (May 2019), pp. 561–566.

[88] Peter Krusche et al. "Best practices for benchmarking germline small-variant calls in human genomes". In: *Nature Biotechnology* (2019).

[89] Aaron McKenna et al. "The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome Research* (2010).

[90] Karen Eilbeck, Aaron Quinlan, and Mark Yandell. "Settling the score: Variant prioritization and Mendelian disease". In: *Nature Reviews Genetics* 18.10 (Oct. 2017), pp. 599–612. URL: http://dx.doi.org/10.1038/nrg.2017.52.

[91] Sue Richards et al. "ACMG Standards and Guidelines Standards and guidelines for the interpretation of sequence variants : a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology". In: 17.January (2015), pp. 1–20.

[92] Colleen Caleshu and Euan A. Ashley. "Taming the genome: Towards better genetic test interpretation". In: *Genome Medicine* 8.1 (June 2016), p. 70. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0325-9.

[93] Melissa A. Kelly et al. "Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: Recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel". In: *Genetics in Medicine* 20.3 (Mar. 2018), pp. 351–359.

[94] Nicola Whiffin et al. "Using high-resolution variant frequencies to empower clinical genome interpretation". In: *Genetics in Medicine* (2017).

[95] Steven M. Harrison, Leslie G. Biesecker, and Heidi L. Rehm. "Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines". In: *Current Protocols in Human Genetics* 103.1 (Sept. 2019). URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cphg.93.

[96] Sean V. Tavtigian et al. "Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework". In: *Genetics in Medicine* 20.9 (Sept. 2018), pp. 1054–1060. URL: http://www.ncbi.nlm.nih.gov/pubmed/29300386%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6336098.

[97] Gundula Povysil et al. "Rare-variant collapsing analyses for complex traits: guidelines and applications". In: *Nature Reviews Genetics* (2019).

[98] Casey A. Gifford et al. "Oligogenic inheritance of a human heart disease involving a genetic modifier". In: *Science* 364.6443 (May 2019), pp. 865–870.

[99] Stéphanie Baulac et al. "Evidence for digenic inheritance in a family with both febrile convulsions and temporal lobe epilepsy implicating chromosomes 18qter and 1q25-q31". In: *Annals of Neurology* (2001).

[100] Alejandro A. Schäffer. "Digenic inheritance in medical genetics". In: *Journal of Medical Genetics* (2013).

[101] S. Fauser, M. Munz, and D. Besch. "Further support for digenic inheritance in Bardet-Biedl syndrome." In: *Journal of medical genetics* (2003).

[102] Maria Kousi and Nicholas Katsanis. "Genetic modifiers and oligogenic inheritance". In: *Cold Spring Harbor Perspectives in Medicine* 5.6 (June 2015), pp. 1–22.

[103] N. Katsanis et al. "Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder". In: *Science* (2001).

[104] Jose L. Badano and Nicholas Katsanis. "Beyond mendel: An evolving view of human genetic disease transmission". In: *Nature Reviews Genetics* 3.10 (Oct. 2002), pp. 779–789.

[105] Ingrid E. Scheffer and Samuel F. Berkovic. "Generalized epilepsy with febrile seizures plus. A genetic disorder with heterogeneous clinical phenotypes". In: *Brain* (1997).

[106] Sumantra Chatterjee et al. "Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease". In: *Cell* (2016).

[107] Perundurai S Dhandapany et al. "A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia." eng. In: *Nature genetics* 41.2 (Feb. 2009), pp. 187–191.

[108] Samir Zaidi et al. "De novo mutations in histone-modifying genes in congenital heart disease". In: *Nature* (2013).

[109] Melina Claussnitzer et al. "A brief history of human disease genetics". In: *Nature* 577.7789 (Jan. 2020), pp. 179–189.

[110] Ashish Sarraju and Joshua W. Knowles. "Genetic Testing and Risk Scores: Impact on Familial Hypercholesterolemia". In: *Frontiers in Cardiovascular Medicine* 6 (Jan. 2019), p. 5. URL: http://www.ncbi.nlm.nih.gov/pubmed/30761309%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6361766.

[111] Mahtab Sharifi et al. "Genetic Architecture of Familial Hypercholesterolaemia". In: *Current Cardiology Reports* 19.5 (May 2017), p. 44. URL: http://link.springer.com/10.1007/s11886-017-0848-8.

[112] Børge G. Nordestgaard et al. "Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to prevent coronary heart disease". In: *European Heart Journal* (2013).

[113] Amit V. Khera et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature Genetics* 50.9 (Sept. 2018), pp. 1219–1224. URL: http://www.ncbi.nlm.nih.gov/pubmed/30104762%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6128408.

[114] Mark Trinder, Gordon A. Francis, and Liam R. Brunham. "Association of Monogenic vs Polygenic Hypercholesterolemia With Risk of Atherosclerotic Cardiovascular Disease". In: *JAMA Cardiology* (Feb. 2020). URL: https://jamanetwork.com/journals/jamacardiology/fullarticle/2760785.

[115] Nicholas J Wald and Robert Old. "The illusion of polygenic disease risk prediction." In: *Genetics in medicine : official journal of the American College of Medical Genetics* 21.8 (Aug. 2019), pp. 1705–1707. URL: http://www.ncbi.nlm.nih.gov/pubmed/30635622.

[116] Gerald F. Watts et al. "Familial hypercholesterolaemia: evolving knowledge for designing adaptive models of care". In: *Nature Reviews Cardiology* (Jan. 2020), pp. 1–18.

[117] Miriam S Udler et al. "Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine". In: *Endocrine Reviews* 40.6 (Dec. 2019), pp. 1500–1520.

[118] Eric C. Wooten et al. "Formin homology 2 domain containing 3 variants associated with hypertrophic cardiomyopathy". In: *Circulation: Cardiovascular Genetics* 6.1 (Feb. 2013), pp. 10–18.

[119] Christopher M. Kramer et al. "Hypertrophic Cardiomyopathy Registry: The rationale and design of an international, observational study of hypertrophic cardiomyopathy". In: *American Heart Journal* 170.2 (Aug. 2015), pp. 223–230. URL: http://www.ncbi.nlm.nih.gov/pubmed/26299218%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4548277.

[120] Willem Hendrik Ouwehand et al. "Whole-genome sequencing of rare disease patients in a national healthcare system". In: *bioRxiv* (Jan. 2020), p. 507244. URL: http://dx.doi.org/10.1101/507244.

[121] Kate L. Thomson et al. "Analysis of 51 proposed hypertrophic cardiomyopathy genes from genome sequencing data in sarcomere negative cases has negligible diagnostic yield". In: *Genetics in Medicine* 21.7 (July 2019), pp. 1576–1584.

[122] Jason Flannick et al. "Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls". In: *Nature* (2019).

[123] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* (2018).

[124] Wouter Van Rheenen et al. "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis". In: *Nature Genetics* (2016).

[125] Sebastian Schafer et al. "Titin-truncating variants affect heart function in disease cohorts and the general population". In: *Nature Genetics* 49.1 (Jan. 2017), pp. 46–53.

[126] Come Raczy et al. "Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms". In: *Bioinformatics* (2013).

[127] Christian Fuchsberger et al. "The genetic architecture of type 2 diabetes". In: *Nature* (2016).

[128] A. L. Williams Amy et al. "Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico". In: *Nature* (2014).

[129] Kirk E. Lohmueller et al. "Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes". In: *American Journal of Human Genetics* (2013).

[130] Wenqing Fu et al. "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants". In: *Nature* (2013).

[131] Cristopher V. Van Hout et al. "Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank". In: *bioRxiv* (2019), p. 572347. URL: https://www.biorxiv.org/content/10.1101/572347v1.

[132] Philip Ewels et al. "MultiQC: Summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* (2016).

[133] P J A Cock et al. "The Sanger FASTQ file format for sequences with quality scores". In: *Nucleic Acids Research* (2010).

[134] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* (2011).

[135] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: A flexible trimmer for Illumina sequence data". In: *Bioinformatics* (2014).

[136] Mark A. Depristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature Genetics* (2011).

[137] A K Thomer et al. "Picard Tools". In: *Conference on Human Factors in Computing Systems - Proceedings*. 2016.

[138] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: (Mar. 2013). URL: http://arxiv.org/abs/1303.3997.

[139] Yuan Freudenberg-Hua et al. "Single nucleotide variation analysis in 65 candidate genes for CNS disorders in representative sample of the European population". In: *Genome Research* (2003).

[140] Ingo Ebersberger et al. "Genomewide comparison of DNA sequences between humans and chimpanzees". In: *American Journal of Human Genetics* (2002).

[141] William McLaren et al. "The Ensembl Variant Effect Predictor". In: *Genome Biology* (2016).

[142] Xiaoming Liu et al. "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs". In: *Human Mutation* (2016).

[143] Pablo Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff". In: *Fly* (2012).

[144] Kim Pruitt et al. "The Reference Sequence (RefSeq) Database". In: *The NCBI Handbook* (2002).

[145] T. Hubbard. "The Ensembl genome database project". In: *Nucleic Acids Research* (2002).

[146] Kristin G. Ardlie et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* (2015).

[147] Daniel S. Herman et al. "Truncations of Titin Causing Dilated Cardiomyopathy". In: *New England Journal of Medicine* 366.7 (Feb. 2012), pp. 619–628. URL: http://www.nejm.org/doi/abs/10.1056/NEJMoa1110186.

[148] Angharad M. Roberts et al. "Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease". In: *Science Translational Medicine* (2015).

[149] Yarden Katz et al. “Analysis and design of RNA sequencing experiments for identifying isoform regulation”. In: *Nature Methods* (2010).

[150] Julian P. Venables et al. “Identification of alternative splicing markers for breast cancer”. In: *Cancer Research* (2008).

[151] Thomas W. Winkler et al. “Quality control and conduct of genome-wide association meta-analyses”. In: *Nature Protocols* (2014).

[152] Shane McCarthy et al. “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature Genetics* (2016).

[153] Alkes L. Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature Genetics* (2006).

[154] P. Menozzi, A. Piazza, and L. Cavalli-Sforza. “Synthetic maps of human gene frequencies in Europeans”. In: *Science* (1978).

[155] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. “Demic expansions and human evolution”. In: *Science* (1993).

[156] Carl A. Anderson et al. “Data quality control in genetic case-control association studies”. In: *Nature Protocols* (2010).

[157] W. J. Youden. “Index for rating diagnostic tests”. In: *Cancer* (1950).

[158] David J. Hand and D. Collett. “Modelling Binary Data.” In: *Applied Statistics* (1993).

[159] Gad Abraham, Yixuan Qiu, and Michael Inouye. “FlashPCA2: principal component analysis of Biobank-scale genotype datasets”. In: *Bioinformatics (Oxford, England)* 33.17 (2017), pp. 2776–2778.

[160] Jan Graffelman and Victor Moreno. “The mid p-value in exact tests for Hardy-Weinberg equilibrium”. In: *Statistical Applications in Genetics and Molecular Biology* (2013).

[161] Shaun Purcell et al. “PLINK: A tool set for whole-genome association and population-based linkage analyses”. In: *American Journal of Human Genetics* (2007).

[162] W N Venables and B D Ripley. *Modern Applied Statistics with S Fourth edition by.* 2002.

[163] John M. Chambers and Trevor J. Hastie. *Statistical models in S.* 2017.

[164] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* (1977).

[165] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* (2010).

[166] Matthew P. Conomos, Michael B. Miller, and Timothy A. Thornton. “Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness”. In: *Genetic Epidemiology* (2015).

[167] Jeffrey Staples et al. “PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent”. In: *American Journal of Human Genetics* (2014).

[168] Edwin B. Wilson. “Probable Inference, the Law of Succession, and Statistical Inference”. In: *Journal of the American Statistical Association* (1927).

[169] Takuro Arimura et al. “Cardiac Ankyrin Repeat Protein Gene (ANKRD1) Mutations in Hypertrophic Cardiomyopathy”. In: *Journal of the American College of Cardiology* (2009).

[170] Adriana Osio et al. “Myozenin 2 is a novel gene for human hypertrophic cardiomyopathy”. In: *Circulation Research* (2007).

[171] Andrew P. Landstrom et al. “Molecular and functional characterization of novel hypertrophic cardiomyopathy susceptibility mutations in TNNC1-encoded troponin C”. In: *Journal of Molecular and Cellular Cardiology* (2008).

[172] Iacopo Olivotto et al. “Gender-related differences in the clinical presentation and outcome of hypertrophic cardiomyopathy”. In: *Journal of the American College of Cardiology* (2005).

[173] Jeffrey B. Geske et al. “Women with hypertrophic cardiomyopathy have worse survival”. In: *European Heart Journal* (2017).

[174] Ethan J. Rowin et al. “Impact of Sex on Clinical Course and Survival in the Contemporary Treatment Era for Hypertrophic Cardiomyopathy”. In: *Journal of the American Heart Association* (2019).

[175] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. “The Missing Diversity in Human Genetic Studies”. In: *Cell* (2019).

[176] Dana Fourey et al. “Prevalence and Clinical Implication of Double Mutations in Hypertrophic Cardiomyopathy”. In: *Circulation: Cardiovascular Genetics* (2017).

[177] Samantha Barratt Ross et al. “Burden of Recurrent and Ancestral Mutations in Families with Hypertrophic Cardiomyopathy”. In: *Circulation: Cardiovascular Genetics* 10.3 (June 2017).

[178] I. Christiaans et al. “Founder mutations in hypertrophic cardiomyopathy patients in the Netherlands”. In: *Neth Heart J* 18.5 (2010), pp. 248–54.

[179] Adam J. Saltzman et al. “Short communication: The cardiac myosin binding protein C Arg502Trp mutation: A common cause of hypertrophic cardiomyopathy”. In: *Circulation Research* 106.9 (May 2010), pp. 1549–1552.

[180] Andrew R Harper et al. “A Re-evaluation of the South Asian MYBPC3Δ25 Intronic Deletion in Hypertrophic Cardiomyopathy.” In: *Circulation. Genomic and precision medicine* (Mar. 2020). URL: `http://www.ncbi.nlm.nih.gov/pubmed/32163302`.

[181] Marielle Alders et al. “The 2373insG mutation in the MYBPC3 gene is a founder mutation, which accounts for nearly one-fourth of the HCM cases in the Netherlands”. In: *European Heart Journal* (2003).

[182] J. C. Barrett et al. “Haploview: Analysis and visualization of LD and haplotype maps”. In: *Bioinformatics* (2005).

[183] J. K. Pritchard and M. Przeworski. “Linkage disequilibrium in humans: Models and data”. In: *American Journal of Human Genetics* (2001).

[184] Berglind Adalsteinsdottir et al. “Nationwide study on hypertrophic cardiomyopathy in iceland evidence of a MYBPC3 founder mutation”. In: *Circulation* (2014).

[185] Xuexia Wang et al. “Adjustment for local ancestry in genetic association analysis of admixed populations”. In: *Bioinformatics* (2011).

[186] Giovanni Montana and Jonathan K. Pritchard. “Statistical tests for admixture mapping with case-control and cases-only data”. In: *American Journal of Human Genetics* (2004).

[187] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. “Negative Controls: A tool for detecting confounding and bias in observational studies”. In: *Epidemiology* (2010).

[188] Benjamin M. Neale et al. “Testing for an unusual distribution of rare variants”. In: *PLoS Genetics* (2011).

[189] Elizabeth Vafiadaki, Demetrios A. Arvanitis, and Despina Sanoudou. “Muscle LIM Protein: Master regulator of cardiac and skeletal muscle functions”. In: *Gene* (2015).

[190] Gwen Freyd, Stuart K. Kim, and H. Robert Horvitz. “Novel cysteine-rich motif and homeodomain in the product of the Caenorhabditis elegans cell lineage gene lin-II”. In: *Nature* (1990).

[191] Olof Karlsson et al. “Insulin gene enhancer binding protein Isl-1 is a member of a novel class of proteins containing both a homeo-and a Cys-His domain”. In: *Nature* (1990).

[192] Jeffrey C. Way and Martin Chalfie. “mec-3, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in C. elegans”. In: *Cell* (1988).

[193] Julie L. Kadrmas and Mary C. Beckerle. “The LIM domain: From the cytoskeleton to the nucleus”. In: *Nature Reviews Molecular Cell Biology* 5.11 (2004), pp. 920–931.

[194] Mehroz Ehsan et al. “Mutant Muscle LIM Protein C58G causes cardiomyopathy through protein depletion”. In: *Journal of Molecular and Cellular Cardiology* (2018).

[195] Alexandre Janin et al. “First identification of homozygous truncating CSRP3 variants in two unrelated cases with hypertrophic cardiomyopathy.” In: *Gene* 676 (Nov. 2018), pp. 110–116. URL: `http://www.ncbi.nlm.nih.gov/pubmed/30012424`.

[196] Martina Lipari et al. “Identification of a variant hotspot in MYBPC3 and of a novel CSRP3 autosomal recessive alteration in a cohort of Polish patients with hypertrophic cardiomyopathy”. In: *Polish Archives of Internal Medicine* 130.2 (Feb. 2020), pp. 89–99. URL: `http://www.ncbi.nlm.nih.gov/pubmed/31919335`.

[197] Xiaowei Li et al. “MLP-deficient human pluripotent stem cell derived cardiomyocytes develop hypertrophic cardiomyopathy and heart failure phenotypes due to abnormal calcium handling”. In: *Cell Death and Disease* 10.8 (Aug. 2019), pp. 1–15.

[198] Hai Fang et al. “A genetics-led approach defines the drug target landscape of 30 immune-related traits”. In: *Nature Genetics* (2019).

[199] Fernando Domínguez et al. “Dilated Cardiomyopathy Due to BLC2-Associated Athanogene 3 (BAG3) Mutations”. In: *Journal of the American College of Cardiology* (2018).

[200] Eric Villard et al. “A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy”. In: *European Heart Journal* (2011).

[201] Ulrike Esslinger et al. “Exome-wide association study reveals novel susceptibility genes to sporadic dilated cardiomyopathy”. In: *PLoS ONE* (2017).

[202] Rocio Toro et al. “Familial dilated cardiomyopathy caused by a novel frameshift in the BAG3 gene”. In: *PLoS ONE* (2016).

[203] Muhammad Arshad Rafiq et al. “Whole exome sequencing identified 1 base pair novel deletion in BCL2-associated athanogene 3 (BAG3) gene associated with severe dilated cardiomyopathy (DCM) requiring heart transplant in multiple family members”. In: *American Journal of Medical Genetics, Part A* (2017).

[204] Xi Fang et al. “Loss-of-function mutations in co-chaperone BAG3 destabilize small HSPs and cause cardiomyopathy”. In: *Journal of Clinical Investigation* (2017).

[205] Sachiko Homma et al. “BAG3 deficiency results in fulminant myopathy and early lethality”. In: *American Journal of Pathology* 169.3 (2006), pp. 761–773.

[206] Christina Klimek et al. “BAG3-mediated proteostasis at a glance”. In: *Journal of Cell Science* 130.17 (Sept. 2017), pp. 2781–2788.

[207] Christian Behl. “Breaking BAG: The Co-Chaperone BAG3 in Health and Disease”. In: *Trends in Pharmacological Sciences* (2016).

[208] Valerie D. Myers et al. “The Multifunctional Protein BAG3: A Novel Therapeutic Target in Cardiovascular Disease”. In: *JACC: Basic to Translational Science* 3.1 (Feb. 2018), pp. 122–131.

[209] Mahmooda Mirza et al. “Dilated cardiomyopathy mutations in three thin filament regulatory proteins result in a common functional phenotype”. In: *Journal of Biological Chemistry* (2005).

[210] Tijana Knezevic et al. “Adeno-Associated Virus Serotype 9–Driven Expression of BAG3 Improves Left Ventricular Function in Murine Hearts With Left Ventricular Dysfunction Secondary to a Myocardial Infarction”. In: *JACC: Basic to Translational Science* (2016).

[211] Harvey S. Hahn et al. “Protein Kinase Cα Negatively Regulates Systolic and Diastolic Function in Pathological Hypertrophy”. In: *Circulation Research* (2003).

[212] Julian C. Braz et al. “PKC-α regulates cardiac contractility and propensity toward heart failure”. In: *Nature Medicine* 10.3 (Mar. 2004), pp. 248–254.

[213] Stephan Lange et al. “MLP and CARP are linked to chronic PKCα signalling in dilated cardiomyopathy”. In: *Nature Communications* (2016).

[214] Ray Hu et al. “Genetic Reduction in Left Ventricular Protein Kinase C-α and Adverse Ventricular Remodeling in Human Subjects”. In: *Circulation. Genomic and precision medicine* (2018).

[215] Dieter O. Fürst et al. “Filamin C-related myopathies: Pathology and mechanisms”. In: *Acta Neuropathologica* 125.1 (Jan. 2013), pp. 33–46.

[216] Juan Gómez et al. “Screening of the Filamin C Gene in a Large Cohort of Hypertrophic Cardiomyopathy Patients”. In: *Circulation: Cardiovascular Genetics* (2017).

[217] Martín F. Ortiz-Genga et al. “Truncating FLNC Mutations Are Associated With High-Risk Dilated and Arrhythmogenic Cardiomyopathies”. In: *Journal of the American College of Cardiology* (2016).

[218] Jessica R. Golbus et al. “Targeted analysis of whole genome sequence data to diagnose genetic cardiomyopathy”. In: *Circulation: Cardiovascular Genetics* (2014).

[219] Rahul C. Deo et al. “Prioritizing causal disease genes using unbiased genomic features”. In: *Genome biology* (2014).

[220] Rene L. Begay et al. “FLNC Gene Splice Mutations Cause Dilated Cardiomyopathy”. In: *JACC: Basic to Translational Science* (2016).

[221] James P Pirruccello et al. “Prevalence and clinical importance of titin truncating variants in adults without known congestive heart failure”. In: *medRxiv* (2019).

[222] Oyediran Akinrinade et al. “Relevance of Titin Missense and Non-Frameshifting Insertions/Deletions Variants in Dilated Cardiomyopathy”. In: *Scientific Reports* 9.1 (Dec. 2019).

[223] Diana Mandelker et al. “Navigating highly homologous genes in a molecular diagnostic setting: A resource for clinical next-generation sequencing”. In: *Genetics in Medicine* (2016).

[224] Sahar Gelfman et al. “Annotating pathogenic non-coding variants in genic regions”. In: *Nature Communications* (2017).

[225] Kishore Jaganathan et al. “Predicting Splicing from Primary Sequence with Deep Learning”. In: *Cell* (2019).

[226] Manatsu Satoh et al. “Structural analysis of the titin gene in hypertrophic cardiomyopathy: Identification of a novel disease gene”. In: *Biochemical and Biophysical Research Communications* 262.2 (Aug. 1999), pp. 411–417.

[227] Nay Aung et al. “Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated with Cardiac Morphogenesis and Heart Failure Development”. In: *Circulation* (Sept. 2019).

[228] R. Michael Sivley et al. “Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures”. In: *American Journal of Human Genetics* 102.3 (Mar. 2018), pp. 415–426.

[229] David S. Lawrie et al. “Strong Purifying Selection at Synonymous Sites in D. melanogaster”. In: *PLoS Genetics* (2013).

[230] Patrick Brest et al. “A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn’s disease”. In: *Nature Genetics* (2011).

[231] Gitte Hoffmann Bruun, Thomas Koed Doktor, and Brage Storstein Andresen. "A synonymous polymorphic variation in ACADM exon 11 affects splicing efficiency and may affect fatty acid oxidation". In: *Molecular Genetics and Metabolism* (2013).

[232] Valeria Pecce et al. "A synonymous RET substitution enhances the oncogenic effect of an in-cis missense mutation by increasing constitutive splicing efficiency". In: *PLoS Genetics* (2018).

[233] Zuben E. Sauna and Chava Kimchi-Sarfaty. "Understanding the contribution of synonymous mutations to human disease". In: *Nature Reviews Genetics* 12 (2011), pp. 683–691.

[234] Gavin Hanson and Jeff Coller. "Translation and Protein Quality Control: Codon optimality, bias and usage in translation and mRNA decay". In: *Nature Reviews Molecular Cell Biology* 19.1 (Jan. 2018), pp. 20–30.

[235] Nicholas T. Ingolia, Jeffrey A. Hussmann, and Jonathan S. Weissman. "Ribosome profiling: Global views of translation". In: *Cold Spring Harbor Perspectives in Biology* 11.5 (May 2019).

[236] Andrew B. Stergachis et al. "Exonic transcription factor binding directs codon choice and affects protein evolution". In: *Science* (2013).

[237] Paul M. Sharp, Therese M.F. Tuohy, and Krzysztof R. Mosurski. "Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes". In: *Nucleic Acids Research* (1986).

[238] Aravinda Chakravarti and Tychele N Turner. "Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families." In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 38.6 (2016), pp. 578–86. URL: `http://www.ncbi.nlm.nih.gov/pubmed/27062178`.

[239] Nicholas J. Schork et al. "Common vs. rare allele hypotheses for complex diseases". In: *Current Opinion in Genetics and Development* 19.3 (June 2009), pp. 212–219.

[240] Sofia Papadimitriou et al. "Predicting disease-causing variant combinations". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.24 (June 2019), pp. 11878–11887.

[241] Paal Skytt Andersen et al. "Diagnostic yield, interpretation, and clinical utility of mutation screening of sarcomere encoding genes in Danish hypertrophic cardiomyopathy patients and relatives". In: *Human Mutation* (2009).

[242] Iacopo Olivotto et al. "Myofilament protein gene mutation screening and outcome of patients with hypertrophic cardiomyopathy". In: *Mayo Clinic Proceedings* (2008).

[243] Murali D Bashyam et al. "A low prevalence of MYH7/MYBPC3 mutations among Familial Hypertrophic Cardiomyopathy patients in India". In: *Molecular and Cellular Biochemistry* 360.1-2 (2012), pp. 373–382.

[244] Diederik W D Kuster and Sakthivel Sadayappan. "MYBPC3's alternate ending: Consequences and therapeutic implications of a highly prevalent 25 bp deletion mutation". In: *Pflugers Archiv European Journal of Physiology* 466.2 (2014), pp. 207–213.

[245] Anshika Srivastava et al. "Association of 25 bp deletion in MYBPC3 gene with left ventricle dysfunction in coronary artery disease patients". In: *PLoS ONE* 6.9 (2011), pp. 1–7.

[246] Surendra Kumar et al. "Role of common sarcomeric gene polymorphisms in genetic susceptibility to left ventricular dysfunction." eng. In: *Journal of genetics* 95.2 (June 2016), pp. 263–272.

[247] David Y Barefield et al. "High-Throughput Diagnostic Assay for a Highly Prevalent Cardiomyopathy-Associated MYBPC3 Variant." eng. In: *Journal of molecular biomarkers & diagnosis* 7.6 (Nov. 2016).

[248] Shiv Kumar Viswanathan et al. "Association of Cardiomyopathy With MYBPC3 D389V and MYBPC3 Δ25bp Intronic Deletion in South Asian Descendants". eng. In: *JAMA Cardiology* 3.6 (June 2018), p. 481.

[249] Paul Robinson et al. "Dilated and hypertrophic cardiomyopathy mutations in troponin and α-tropomyosin have opposing effects on the calcium affinity of cardiac thin filaments". In: *Circulation Research* 101.12 (2007), pp. 1266–1273.

[250] Edward P Debold et al. "Hypertrophic and dilated cardiomyopathy mutations differentially affect the molecular force generation of mouse alpha-cardiac myosin in the laser trap assay." eng. In: *American journal of physiology. Heart and circulatory physiology* 293.1 (July 2007), pp. 284–91.

[251] Raquel Yotti, Christine E Seidman, and Jonathan G Seidman. "Advances in the Genetic Basis and Pathogenesis of Sarcomere Cardiomyopathies." eng. In: *Annual review of genomics and human genetics* (Apr. 2019).

[252] Stephan Waldmüller et al. "Novel deletions in MYH7 and MYBPC3 identified in Indian families with familial hypertrophic cardiomyopathy". In: *Journal of Molecular and Cellular Cardiology* (2003).

[253] Samuel P. Dickson et al. "Rare Variants Create Synthetic Genome-Wide Associations". In: *PLoS Biology* (2010).

[254] Naomi R. Wray, Shaun M. Purcell, and Peter M. Visscher. "Synthetic associations created by rare variants do not explain most GWAS results". In: *PLoS Biology* (2011).

[255] Gisela Orozco, Jeffrey C. Barrett, and Eleftheria Zeggini. "Synthetic associations in the context of genome-wide association scan signals". In: *Human Molecular Genetics* (2010).

[256] Eric Vallabh Minikel et al. "Quantifying prion disease penetrance using large population control cohorts". In: *Science Translational Medicine* 8.322 (Jan. 2016), pp. 9–322. URL: `http://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aad5169`.

[257] Evangelos Evangelou et al. "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits". In: *Nature Genetics* (2018).

[258] Robert J. Klein et al. "Complement factor H polymorphism in age-related macular degeneration". In: *Science* 308.5720 (Apr. 2005), pp. 385–389.

[259] Sergey Nejentsev et al. "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A". In: *Nature* 450.7171 (Dec. 2007), pp. 887–892.

[260] Thorsteinn Bjornsson et al. “A rare missense mutation in MYH6 associates with non-syndromic coarctation of the aorta”. In: *European Heart Journal* 39.34 (Sept. 2018), pp. 3243–3249. URL: https://academic.oup.com/eurheartj/article/39/34/3243/4953519.

[261] Caroline F. Wright et al. “Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting”. In: *American Journal of Human Genetics* 104.2 (2019), pp. 275–286.

[262] Joël Zlotogora. “Penetrance and expressivity in the molecular age”. In: *Genetics in Medicine* 5 (2003), pp. 347–352.

[263] Stephen P. Page et al. “Cardiac myosin binding protein-C mutations in families with hypertrophic cardiomyopathy: Disease expression in relation to age, gender, and long term outcome”. In: *Circulation: Cardiovascular Genetics* (2012).

[264] Jérôme Carayol and Catherine Bonaïti-Pellié. “Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset”. In: *Genetic Epidemiology* (2004).

[265] Christopher M. Haggerty et al. “Electronic health record phenotype in subjects with genetic variants associated with arrhythmogenic right ventricular cardiomyopathy: A study of 30,716 subjects with exome sequencing”. In: *Genetics in Medicine* 19.11 (Nov. 2017), pp. 1245–1252.

[266] Robert C. Green et al. “ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing”. In: *Genetics in Medicine* (2013).

[267] Sarah S. Kalia et al. “Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics”. In: *Genetics in Medicine* (2017).

[268] Carla G. Van El et al. “Whole-genome sequencing in health care”. In: *European Journal of Human Genetics* (2013).

[269] *Findings | Genomics England*. URL: https://www.genomicsengland.co.uk/information-for-participants/findings/.

[270] Spiros C. Denaxas et al. “Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)”. In: *International Journal of Epidemiology* (2012).

[271] “UK Biobank-Exome Data Release FAQs”. In: (2019). URL: https://www.biorxiv.org/content/10.1101/572347v1.

[272] Allison A. Regier et al. “Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects”. In: *Nature Communications* 9.1 (Dec. 2018), p. 4038. URL: http://www.nature.com/articles/s41467-018-06159-4.

[273] Tongqiu Jia et al. “Thousands of missing variants in the UK BioBank are recoverable by genome realignment”. In: *bioRxiv* (Dec. 2019), p. 868570.

[274] Joel N. Buxbaum and Frederick L. Ruberg. "Transthyretin V122I (pV142I) cardiac amyloidosis: An age-dependent autosomal dominant cardiomyopathy too common to be overlooked as a cause of significant heart disease in elderly African Americans". In: *Genetics in Medicine* (2017).

[275] Matthew W. Knuiman et al. "Adjustment for regression dilution in epidemiological regression analyses". In: *Annals of Epidemiology* (1998).

[276] V. Moskvina et al. "Design of case-controls studies with unscreened controls". In: *Annals of Human Genetics* (2005).

[277] Jeffrey A. Towbin et al. "2019 HRS expert consensus statement on evaluation, risk stratification, and management of arrhythmogenic cardiomyopathy". In: *Heart Rhythm* (2019).

[278] Laura M. Beskow et al. "Recommendations for ethical approaches to genotype-driven research recruitment". In: *Human Genetics* (2012).

[279] Alexander G. Bick et al. "Burden of rare sarcomere gene variants in the framingham and jackson heart study cohorts". In: *American Journal of Human Genetics* 91.3 (Sept. 2012), pp. 513–519.

[280] Amit V. Khera et al. "Rare Genetic Variants Associated With Sudden Cardiac Death in Adults". In: *Journal of the American College of Cardiology* 74.21 (Nov. 2019), pp. 2623–2634.

[281] Andrew R. Harper, Shalini Nayee, and Eric J. Topol. "Protective alleles and modifier variants in human health and disease". In: *Nature Reviews Genetics* 16.12 (Dec. 2015), pp. 689–701.

[282] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1 (July 2017), pp. 5–22.

[283] Mari E.K. Niemi et al. "Common genetic variants contribute to risk of rare severe neurodevelopmental disorders". In: *Nature* 562.7726 (Oct. 2018), pp. 268–271.

[284] Karoline B Kuchenbaecker et al. "Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers". In: *JNCI: Journal of the National Cancer Institute* 109.7 (July 2017). URL: `https://academic.oup.com/jnci/article/doi/10.1093/jnci/djw302/3064534`.

[285] Akl C Fahed et al. "Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer". In: (2019). URL: `http://dx.doi.org/10.1101/19013086`.

[286] Fergus J. Couch et al. "Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk". In: *PLoS Genetics* 9.3 (Mar. 2013). Ed. by Kent W. Hunter, e1003212. URL: `https://dx.plos.org/10.1371/journal.pgen.1003212`.

[287] A.R. Harper and E.J. Topol. "Pharmacogenomics in clinical practice and drug development". In: *Nature Biotechnology* 30.11 (2012).

[288] Andrew D. Skol et al. "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies". In: *Nature Genetics* (2006).

[289] Guillaume Lettre, Christoph Lange, and Joel N. Hirschhorn. "Genetic model testing and statistical power in population-based association studies of quantitative traits". In: *Genetic Epidemiology* (2007).

[290] Itsik Pe'er et al. "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants". In: *Genetic Epidemiology* (2008).

[291] John W. Belmont et al. "A haplotype map of the human genome". In: *Nature* (2005).

[292] Sayantan Das et al. "Next-generation genotype imputation service and methods". In: *Nature Genetics* (2016).

[293] Po Ru Loh, Pier Francesco Palamara, and Alkes L. Price. "Fast and accurate long-range phasing in a UK Biobank cohort". In: *Nature Genetics* (2016).

[294] Matti Pirinen, Peter Donnelly, and Chris C.A. Spencer. "Including known covariates can reduce power to detect genetic effects in case-control studies". In: *Nature Genetics* (2012).

[295] Wei Zhou et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". In: *Nature Genetics* (2018).

[296] Daniel John Lawson et al. "Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?" In: *Human Genetics* (2020).

[297] Christopher C. Chang et al. "Second-generation PLINK: Rising to the challenge of larger and richer datasets". In: *GigaScience* (2015).

[298] Tinca J.C. Polderman et al. "Meta-analysis of the heritability of human traits based on fifty years of twin studies". In: *Nature Genetics* 47.7 (June 2015), pp. 702–709.

[299] Jian Yang et al. "GCTA: A tool for genome-wide complex trait analysis". In: *American Journal of Human Genetics* (2011).

[300] Jian Yang et al. "Concepts, estimation and interpretation of SNP-based heritability". In: *Nature Genetics* (2017).

[301] Kangcheng Hou et al. "Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture". In: *Nature Genetics* 51.8 (Aug. 2019), pp. 1244–1251.

[302] S. Zdravkovic et al. "Heritability of death from coronary heart disease: A 36-year follow-up of 20 966 Swedish twins". In: *Journal of Internal Medicine* 252.3 (2002), pp. 247–254.

[303] David G. Clayton et al. "Population structure, differential bias and genomic control in a large-scale, case-control association study". In: *Nature Genetics* 37.11 (Nov. 2005), pp. 1243–1246.

[304] B. Devlin, Kathryn Roeder, and Larry Wasserman. "Genomic control, a new approach to genetic-based association studies". In: *Theoretical Population Biology* 60.3 (2001), pp. 155–166.

[305] Paul R. Burton et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (June 2007), pp. 661–678.

[306] Loic Yengo et al. “Meta-analysis of genome-wide association studies for height and body mass index in ˜700 000 individuals of European ancestry”. In: *Human Molecular Genetics* 27.20 (2018), pp. 3641–3649.

[307] Brendan Bulik-Sullivan et al. “LD score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature Genetics* 47.3 (Feb. 2015), pp. 291–295.

[308] Yang Wu et al. “Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data”. In: *Genome Biology* (2017).

[309] Sara L. Pulit, Sera A.J. de With, and Paul I.W. de Bakker. “Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations”. In: *Genetic Epidemiology* (2017).

[310] Randall C. Johnson et al. “Accounting for multiple comparisons in a genome-wide association study (GWAS)”. In: *BMC Genomics* 11.1 (Dec. 2010), p. 724.

[311] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300. URL: http://doi.wiley.com/10.1111/j.2517-6161.1995.tb02031.x.

[312] Christopher P. Nelson et al. “Association analyses based on false discovery rate implicate new loci for coronary artery disease”. In: *Nature Genetics* (2017).

[313] John D. Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (Aug. 2002), pp. 479–498. URL: http://doi.wiley.com/10.1111/1467-9868.00346.

[314] John D. Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (Aug. 2003), pp. 9440–9445.

[315] John D Storey. *False Discovery Rates Multiple Hypothesis Testing*. Tech. rep. January. 2010, pp. 1–7. URL: http://genomine.org/papers/Storey_FDR_2011.pdf.

[316] Jian Yang et al. “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits”. In: *Nature Genetics* 44.4 (Apr. 2012), pp. 369–375. URL: http://www.nature.com/articles/ng.2213.

[317] Reedik Mägi and Andrew P. Morris. “GWAMA: Software for genome-wide association meta-analysis”. In: *BMC Bioinformatics* (2010).

[318] Hilary K. Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature Genetics* (2015).

[319] Eddie Cano-Gamez and Gosia Trynka. “From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases”. In: *Frontiers in Genetics* 11 (May 2020), p. 424. URL: www.frontiersin.org.

[320] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature Reviews Genetics* 19.8 (Aug. 2018), pp. 491–504. URL: www.nature.com/nrg.

[321] François Aguet et al. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *bioRxiv* (2019).

[322] Anthony D. Schmitt et al. “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome”. In: *Cell Reports* (2016).

[323] Zagaa Odgerel et al. “Inheritance patterns and phenotypic features of myofibrillar myopathy associated with a BAG3 mutation”. In: *Neuromuscular Disorders* 20.7 (July 2010), pp. 438–442.

[324] Duygu Selcen et al. “Mutation in BAG3 causes severe dominant childhood muscular dystrophy”. In: *Annals of Neurology* 65.1 (Jan. 2009), pp. 83–89.

[325] Benjamin B. Sun et al. “Genomic atlas of the human plasma proteome”. In: *Nature* 558.7708 (June 2018), pp. 73–79.

[326] *10_119656173_G_A | Open Targets Genetics*. URL: `https://genetics.opentargets.org/variant/10_119656173_G_A`.

[327] Tijana Knezevic et al. “BAG3: a new player in the heart failure paradigm”. In: *Heart Failure Reviews* 20.4 (July 2015), pp. 423–434.

[328] Sonia Shah et al. “Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure”. In: *Nature Communications* 11.163 (2020).

[329] Pim van der Harst et al. “52 Genetic Loci Influencing Myocardial Mass”. In: *Journal of the American College of Cardiology* 68.13 (Sept. 2016), pp. 1435–1448.

[330] Tomoki Ushijima et al. “The actin-organizing formin protein Fhod3 is required for postnatal development and functional maintenance of the adult heart in mice”. In: *Journal of Biological Chemistry* 293.1 (2018), pp. 148–162.

[331] Sho Matsuyama et al. “Interaction between cardiac myosin-binding protein C and formin Fhod3”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.19 (May 2018), E4386–E4395.

[332] Thomas P. Cappola et al. “Loss-of-function DNA sequence variant in the CLCNKA chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.6 (Feb. 2011), pp. 2456–2461.

[333] Liang Yi Juo et al. “HSPB7 interacts with dimerized FLNC and its absence results in progressive myopathy in skeletal muscles”. In: *Journal of Cell Science* 129.8 (Apr. 2016), pp. 1661–1670.

[334] Honghuang Lin et al. “Common and Rare Coding Genetic Variation Underlying the Electrocardiographic PR Interval”. In: *Circulation. Genomic and precision medicine* 11.5 (May 2018), e002037.

[335] Stuart J. Smith et al. “The cardiac-restricted protein ADP-ribosylhydrolase-like 1 is essential for heart chamber outgrowth and acts on muscle actin filament assembly”. In: *Developmental Biology* 416.2 (Aug. 2016), pp. 373–388.

[336] P. D. Pion et al. “Myocardial failure in cats associated with low plasma taurine: A reversible cardiomyopathy”. In: *Science* (1987).

[337] N. Sydney Moise et al. “Dietary taurine deficiency and dilated cardiomyopathy in the fox”. In: *American Heart Journal* (1991).

[338] Keiko Takihara et al. “Beneficial effect of taurine in rabbits with chronic congestive heart failure”. In: *American Heart Journal* (1986).

[339] Joanna L. Kaplan et al. “Taurine deficiency and dilated cardiomyopathy in golden retrievers fed commercial diets”. In: *PLoS ONE* (2018).

[340] Stefan Neubauer et al. “Distinct Subgroups in Hypertrophic Cardiomyopathy in the NHLBI HCM Registry”. In: *Journal of the American College of Cardiology* 74.19 (Nov. 2019), pp. 2333–2345. URL: https://linkinghub.elsevier.com/retrieve/pii/S0735109719376831.

[341] Joseph K. Pickrell et al. “Detection and interpretation of shared genetic influences on 42 human traits”. In: *Nature Genetics* 48.7 (July 2016), pp. 709–717.

[342] FinnGen. *FinnGen: Cardiomyopathy, Hypertrophic obstructive*. 2020. URL: https://r2.finngen.fi/pheno/I9_CARDMYOHYP.

[343] Mridula Nambiar and Gerald R. Smith. “Repression of harmful meiotic recombination in centromeric regions”. In: *Seminars in Cell and Developmental Biology* 54 (June 2016), pp. 188–197.

[344] Alkes L. Price et al. “Long-Range LD Can Confound Genome Scans in Admixed Populations”. In: *American Journal of Human Genetics* 83.1 (July 2008), pp. 132–135.

[345] Michael E. Weale. “Quality control for genome-wide association studies”. In: *Methods in Molecular Biology* 628 (2010), pp. 341–372.

[346] *Regions of high linkage disequilibrium (LD) - Genome Analysis Wiki*. URL: https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD).

[347] Mihir A. Kamat et al. “PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations”. In: *Bioinformatics (Oxford, England)* (2019).

[348] H. H.H. Göring, J. D. Terwilliger, and J. Blangero. “Large upward bias in estimation of locus-specific effects from genomewide scans”. In: *American Journal of Human Genetics* 69.6 (2001), pp. 1357–1369. URL: https://pubmed.ncbi.nlm.nih.gov/11593451/.

[349] Carlo Fumagalli et al. “Association of Obesity with Adverse Long-term Outcomes in Hypertrophic Cardiomyopathy”. In: *JAMA Cardiology* (2019).

[350] George D. Smith and Shah Ebrahim. “’Mendelian randomization’: Can genetic epidemiology contribute to understanding environmental determinants of disease?” In: *International Journal of Epidemiology* 32.1 (2003), pp. 1–22.

[351] Gibran Hemani et al. “The MR-base platform supports systematic causal inference across the human phenome”. In: *eLife* (2018).

[352] Alexandra Dainis et al. “Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3.” In: *Circulation. Genomic and precision medicine* 12.5 (May 2019), e002464. URL: http://www.ncbi.nlm.nih.gov/pubmed/31112421.

[353] Adam F. Johnson, Ha T. Nguyen, and Reiner A. Veitia. “Causes and effects of haploinsufficiency”. In: *Biological Reviews* 94.5 (Oct. 2019), pp. 1774–1785. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12527.

[354] *Definition of haploinsufficiency - NCI Dictionary of Genetics Terms - National Cancer Institute*. URL: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/haploinsufficiency.

[355] Zornitza Stark et al. "Integrating Genomics into Healthcare: A Global Responsibility". In: *American Journal of Human Genetics* 104.1 (Jan. 2019), pp. 13–20.

[356] Sara L. Pulit et al. "Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry". In: *Human Molecular Genetics* (2019).

[357] Joseph K. Pickrell. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits". In: *American Journal of Human Genetics* (2014).